UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

DESARROLLO DE UNA APLICACIÓN PARA LA CONSULTA Y ADMINISTRACIÓN DE UN CORPUS LINGÜÍSTICO ELECTRÓNICO. UNA APORTACIÓN TECNOLÓGICA AL CORPUS HISTÓRICO DEL ESPAÑOL EN MÉXICO.

TESIS PROFESIONAL

MANUEL ALEJANDRO GÓMEZ CHAVARRIA

CLAUDIA GONZÁLEZ CASTAÑEDA

MÉXICO, D.F. **2010**

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

DESARROLLO DE UNA APLICACIÓN PARA LA CONSULTA Y ADMINISTRACIÓN DE UN CORPUS LINGÜÍSTICO ELECTRÓNICO. UNA APORTACIÓN TECNOLÓGICA AL CORPUS HISTÓRICO DEL ESPAÑOL EN MÉXICO.

TESIS PROFESIONAL
QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN INFORMÁTICA

PRESENTAN:

MANUEL ALEJANDRO GÓMEZ CHAVARRIA

CLAUDIA GONZÁLEZ CASTAÑEDA

ASESOR:

MTRO. CARLOS FRANCISCO MÉNDEZ CRUZ

MÉXICO, D.F. **2010**

Agradecimientos

Agradecemos especialmente a nuestra máxima casa de estudios, la Universidad Nacional Autónoma de México, por ayudar a forjarnos como hombres y mujeres con valores, comprometidos con el desarrollo y el progreso de nuestro país. También agradecemos al Grupo de Ingeniería Lingüística del Instituto de Ingeniería de la UNAM por proporcionarnos las herramientas para la realización de este trabajo.

Asimismo gracias a nuestro asesor el Mtro. Carlos Francisco Méndez Cruz por invitarnos a formar parte del GIL, así como por su constante apoyo y confianza depositada; por alentar en nosotros el deseo de aprender. Gracias también al Dr. Alfonso Medina por la oportunidad de formar parte del equipo de desarrolló del CHEM, por demostrar su seguridad y fe para con nosotros.

Por último, queremos agradecer a los proyectos DGAPA-PAPIIT, IN 402008, "Glutinometría y variación dialectal" y CONACYT, "Extracción de conocimiento lexicográfico a partir de textos de Internet" por el apoyo económico que nos brindaron durante la realización de esta tesis.

Claudia

A mis padres:

Gracias por hacer este momento de mi vida posible, por ayudarme en el camino que ahora me ha traído hasta aquí. Me siento muy afortunada y orgullosa de ambos.

A mi madre:

Gracias infinitas porque nunca habrá forma de agradecer todo su esfuerzo. Su amor y su apoyo desmedido siempre serán para mí lo más valioso.

A mi padre:

Por su cariño, sus enseñanzas, su constancia y comprensión para conmigo: gracias. Pero sobre todo, por tomar una tarea que no era suya y a pesar de eso cumplirla como el mejor, sin medida y sin esperar nada a cambio.

A mi hermana:

Por ser un ejemplo para mí, por su lealtad, su apoyo y por la fe que siempre ha demostrado tener en mí. También gracias por el cariño y la compañía.

A toda mi familia:

A cada miembro de mi familia, pero muy especialmente a mi abuelita, por ser el pilar de toda mi vida, gracias por cada palabra que me ha enseñado a ser mejor.

A mi novio y compañero de tesis:

Gracias, por permitirme compartir con él esté sueño que también es suyo. Por motivarme a realizar este trabajo y brindarme todo su apoyo, cariño y confianza desde el inicio de este proyecto. Gracias por ser ahora parte de mi vida.

A mis amigas:

A todas ellas gracias, por ser parte de mi vida en algún momento, atesoro grandes recuerdos con cada una.

Al Mtro. Gabriel Guevara Gutiérrez:

Profesor y amigo, gracias por su disposición y sus valiosas enseñanzas no sólo en el ámbito escolar también en lo personal. Porque admiro su labor como docente y su iniciativa de formar profesionistas competitivos pero sobretodo humanos.

Finalmente agradezco a todas aquellas personas que ya no regresaran a mi vida pero que siempre llevaré en mi corazón

.

Manuel Alejandro

A mis padres:

Quiero agradecerles a ustedes por ser el pilar más importante dentro de mi vida, del cual me he apoyado en los momentos más difíciles. Por haberme brindado su total apoyo y confianza. Gracias por guiarme y demostrarme que por muy difícil que sea cumplir una meta siempre es posible alcanzarla trabajando duro y no dejándome caer. Nunca podré agradecerles lo suficiente por todo lo que han hecho por mí.

A mis hermanos:

A ustedes Roberto, Nelsayet y Yareth por crecer, convivir y jugar conmigo desde que nací, les agradezco tanto los buenos como los malos momentos que hemos compartido, guardo maravillosos recuerdos.

A mi abuelita:

Por todo el cariño que me has brindado durante toda mi vida.

A toda mi familia:

Sin ser menos importantes, a todos los demás integrantes de mi familia. Siempre he pensado que no es necesario un lazo de sangre para considerar a alguien que aprecio como parte de mi familia, por eso gracias a mis padrinos.

A mi novia Claudia:

Porque sin ti no hubiese llegado al lugar donde me encuentro actualmente. Te agradezco todos los momentos que hemos tenidos juntos. Gracias por hacerme feliz.

A mis profesores:

Al Mtro. Gabriel Guevara Gutiérrez por su disposición y por demostrarme que mi crecimiento personal y académico depende totalmente de mis decisiones.

Al profesor Luis Octavio Ramírez Fernández por brindarme además de sus enseñanzas su amistad.

A mis amigos:

A todos y cada uno de ustedes que me han acompañado durante toda mi vida personal y académica. Porque con ustedes he compartido los mejores momentos de mi vida.

Índice general

Índice de Figuras

Índice de tablas

Introducción

La Universidad Nacional Autónoma de México, preocupada por la cultura e investigación en nuestro país, dedica importantes esfuerzos por organizar y realizar investigaciones que permitan incrementar la cultura en diversos ámbitos. Para lograrlo, la UNAM se ayuda, entre otros recursos, de un gran número de institutos dedicados a la investigación en distintas áreas del conocimiento.

Entre la variedad de institutos dedicados a la investigación en nuestro país se reconoce al Instituto de Ingeniería de la UNAM (IIUNAM) como "el centro de investigación en diversas áreas de la ingeniería más productivo del país" **(WEB 01)**. Este instituto comenzó sus actividades a mediados de la década de los 50 con la única misión de contribuir al desarrollo del país y al bienestar de la sociedad a través de la investigación en ingeniería y de la formación de recursos humanos. Para cumplir con su misión, el IIUNAM actualmente se integra por varios grupos de investigación, cada uno orientado a diferentes áreas de la ingeniería.

Uno de estos grupos es el Grupo de Ingeniería Lingüística (GIL), el cual surge formalmente en agosto de 1999, y está dedicado al estudio de la ingeniería lingüística. El objetivo de este grupo es desarrollar sistemas informáticos, herramientas, tecnología, métodos y productos que permitan resolver problemas sociales mediante el estudio y procesamiento automatizado del lenguaje. Para lo anterior, el GIL está desarrollando proyectos con diferentes líneas de investigación como son: lingüística de corpus, terminología automática, extracción conceptual, entre otras. Para estos proyectos el GIL se ha beneficiado con el apoyo de patrocinadores como el Consejo Nacional de Ciencia y Tecnología (CONACYT) así como de la Dirección General de Apoyo al Personal Académico de la UNAM (DGAPA), a través de su Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT).

Dentro de los proyectos en los que actualmente trabaja el GIL está el enfocado al desarrollo de corpus lingüísticos electrónicos[1]. Mediante el desarrollo de investigación y tecnología, el grupo comenzó a construir el **Corpus Histórico del Español en México (CHEM)** (proyecto DGAPA-PAPIIT IN 400905 "Constitución del corpus histórico del español de México "), el cual tiene como objetivo principal "constituir un corpus diacrónico del español de México, con textos escritos en la Nueva España y el México independiente (siglos XVI al XIX), que pueda ser utilizado por estudiosos del lenguaje mediante una aplicación Web diseñada con herramientas para hacer exploraciones sobre el mismo" **(WEB, 02)**.

El GIL, como mencionamos, ha venido desarrollando el CHEM desde hace varios años, para ser precisos desde el año 2005, y ha logrado poner en línea una primera versión con el resultado de sus investigaciones y desarrollos. Dentro de éstos se distingue una herramienta llamada generador de concordancias, para la consulta de un acervo de alrededor de trescientos documentos electrónicos, todos ellos proporcionados por la Dr. Concepción Company del Instituto de Investigaciones Filológicas.

Sin embargo, aunque el CHEM cumple de alguna forma con el objetivo para el que fue creado, surgió la necesidad de mejorar la primera versión para incluir en ella los avances, desarrollos y documentos electrónicos que se han generado desde que ésta vio la luz. La mejora es necesaria para reparar algunos de los problemas que han surgido de la primera versión y permitir que el CHEM pueda seguir cumpliendo con su objetivo. Además un corpus más grande puede ofrecer resultados más confiables a los estudiosos del lenguaje; y un mayor número de herramientas brinda más opciones para analizar el corpus.

Precisamente en la necesidad anterior es donde tiene cabida esta tesis. Siendo estudiantes de la Licenciatura en Informática de la Facultad de Contaduría y Administración de la Universidad Nacional Autónoma de México (FCA), fuimos invitados a formar parte del proyecto del CHEM por el Dr. Medina y el Mtro. Méndez, ambos integrantes del GIL, quienes

[1] Un corpus lingüístico electrónico es una colección de textos o documentos que permite, mediante el uso de sistemas de información: generar resumen automático de textos, extracción y recuperación de información e incluso la aplicación de minería de textos, entre otros procedimientos. Más adelante dedicaremos algunos capítulos para ampliar este tema.

habiéndonos expuesto en qué consistía el proyecto y cuál era la problemática, dejaron a nuestra consideración el aceptar o no la invitación. Evidentemente aceptamos.

Entonces, la finalidad de esta tesis es desarrollar una investigación cuyo resultado sea la construcción de una aplicación web que integre al CHEM varios aspectos. El primero de ellos es un control de usuarios y accesos, es decir, dependiendo del tipo de usuario, esté podrá hacer uso o no de los contenidos de la aplicación, ya sean las herramientas o el mismo corpus. También se contempla la construcción de una nueva herramienta de análisis de corpus lingüísticos: un generador de estadísticas, el cual obtiene precisamente estadísticas sobre la asociación que existe entre las palabras del corpus. Por último, se contempla el desarrollo de una herramienta que permita visualizar los documentos completos.

Derivado de lo anterior, el título de esta tesis es: *Desarrollo de una aplicación para la administración y consulta de un Corpus Lingüístico Electrónico. Una aportación tecnológica al Corpus Histórico del Español en México.* Aclaramos que esta investigación es una aportación dado que se propone la inclusión de tecnologías y métodos para el desarrollo de nuevas herramientas de explotación del corpus, así como para la mejora de las funcionalidades ya existentes. De igual forma recalcamos que es sólo "una aportación", porque después de realizada esta tesis, se continuará desarrollando tecnología para el procesamiento del CHEM.

Las motivaciones que nos llevaron a aceptar formar parte del equipo de desarrollo del CHEM fueron, en primer lugar, el reto que representa el conjugar los conocimientos aprendidos en la licenciatura con un área tan desconocida para nosotros como lo era la lingüística de corpus. Pero, la informática, desde hace ya varias décadas, se ha venido aplicando a diversas áreas del conocimiento. Áreas tan variadas como la aeronáutica, la biología o la genética se han visto beneficiadas por las ventajas que ésta conlleva: rapidez, confianza y oportunidad en la obtención de información, sólo por mencionar algunas. ¿Por qué no la lingüística?

En segundo lugar, otra de las motivaciones es la de demostrar el alcance que puede tener la informática para desarrollar herramientas computacionales útiles en la lingüística de corpus, así como para resolver problemas en cuanto a la gestión de la información de un corpus lingüístico electrónico.

En cuanto a la relevancia de esta investigación, podemos decir que a partir de los resultados de esta tesis será posible desarrollar e integrar, con relativa facilidad, nuevas herramientas para explotación del CHEM, por ejemplo un generador de *colocaciones*[2], un analizador morfológico o incluso un resumidor automático, por mencionar algunas.

Esta investigación también sienta la base para una mejor administración de futuros requerimientos, al considerarse el desarrollo de opciones no sólo para registrar, sino para gestionar nuevas herramientas, permisos o tipos de usuario[3].

0.1 Planteamiento del problema

A continuación describimos brevemente el funcionamiento de la primera versión del CHEM, ya disponible en Web al inicio de esta tesis, con la intención de manifestar algunos de los problemas de la misma[4] .

Comenzamos señalando que la mayoría de los contenidos del CHEM se presentan de manera sencilla y requieren de mayor información[5]. La primera versión del CHEM se concreta a mostrar un menú con opciones para conocer más sobre el mismo, una ventana inicial de bienvenida que permite el registro y acceso usuarios, y una ventana para realizar búsquedas de concordancias.

Para que un usuario pueda utilizar la primera versión del CHEM requiere, en primer lugar, de ir a la ventana de registro del usuario. Posteriormente, a través de un nombre de usuario y una contraseña, se le permitirá utilizar el generador de concordancias, única herramienta de explotación para el corpus hasta esta versión. El registro de usuario

[2] En el capítulo 5 hacemos una breve descripción de lo que son las colocaciones en un corpus lingüístico.

[3] De igual forma, la administración de herramientas, permisos y tipos de usuarios son tratadas con mayor claridad en capítulos posteriores.

[4] Aclaramos que el orden en que serán expuestos los problemas de la primera versión del CHEM no tiene nada que ver con su importancia, lo hacemos de esta forma porque así se realiza normalmente el proceso para utilizar el corpus.

[5] En esta tesis nos referiremos a contenidos cuando hablamos de textos que brindan información acerca del CHEM o aquellos que permiten la navegación en el mismo.

básicamente es un formulario que permite que el usuario proporcione algunos de sus datos personales para que estos puedan ser guardados en una base de datos. Sin embargo, dicho registro no resulta del todo efectivo y veraz, ya que carece de algún método o herramienta que se encargue de asegurar que el usuario está proporcionando todos los datos que se le requieren y además de garantizar que de algún modo esos datos son confiables. Por lo anterior, los dos grandes problemas del registro de usuarios son: en primer lugar, que se consiente que los usuarios omitan ciertos datos y, en segundo lugar, que esos datos pueden ser incorrectos.

Mencionábamos que el usuario requiere registrarse para poder hacer uso del generador de concordancias, regresando a esta herramienta indicamos que a las concordancias se pueden aplicar algunas opciones para su búsqueda y presentación. Para la primera, el generador brinda la opción de buscar por siglo, mientras que para la segunda contiene un espacio para indicar el número de caracteres que se desea que aparezcan en la ventana de la concordancia[6].

De los párrafos anteriores se desprende que los usuarios registrados, una vez que acceden al corpus con su usuario y contraseña, pueden tener acceso a la búsqueda de concordancias sin ninguna restricción, por ejemplo de tipo de usuario. Con "tipo de usuario" nos referimos a que no se contemplaba ninguna distinción entre los usuarios que utilizan los materiales y herramientas propias del corpus, es decir, todos los usuarios pueden procesar todos los documentos sin importar alguna categoría de los mismos. Una restricción es necesaria debido a que los materiales que procesa el CHEM son documentos proporcionados por investigadores y estos tienen permisos inherentes para poder ser procesados o visualizados por terceros[7].

[6] En la sección enfocada a la investigación de concordancias en el capítulo 2 profundizaremos en el tema de la ventana de la concordancia.

[7] Al igual que el tema de la ventana de la concordancia, el tema de los permisos de documentos será tratado en detalle más adelante.

Asimismo, los usuarios que visitan el CHEM dentro de las instalaciones de la UNAM no reciben ningún beneficio, lo que ya ocurre por ejemplo con otros sistemas de la Universidad y que se cree puede ser alguna distinción para la comunidad universitaria[8].

De lo anterior se desprende la necesidad del manejo de diferentes tipos de usuarios. Esta tesis contempla en principio, para la administración del CHEM, la existencia de dos principales tipos: usuarios registrados y usuarios anónimos. Ambos subdivididos a su vez en subtipos para una mejor administración. La idea es asignar privilegios para utilizar las herramientas de explotación del corpus dependiendo del tipo de usuario. Por ejemplo, un usuario registrado podría utilizar la herramienta "generador de concordancias" así como ver algunos de los documentos en su totalidad, mientras que un usuario anónimo fuera de la UNAM sólo podría utilizar el "generador de concordancias"; otro que se encuentre dentro de la UNAM también podría tener la oportunidad de ver algunos de los documentos, pero no todos.

Como mencionábamos, la primera versión únicamente cuenta con la herramienta para generar concordancias. Es precisamente en este punto donde surge otra necesidad ya que dentro del objetivo principal del CHEM está el contar con herramientas de utilidad para investigadores y académicos, entre otros. El problema es el siguiente, aunque se les presenten las concordancias, esta información puede no ser suficiente para su investigación. Por lo anterior, sería conveniente darles la posibilidad de conocer los documentos completos de los que son extraídas las concordancias; esto además implica un mayor reto: la seguridad de los documentos que se publicarán.

Esta tesis contempla dos principales formas para mantener, en la medida de los posible, la seguridad de los documentos: permisos de documento y restricción de copiado e impresión de documento. Es importante que los documentos tengan permisos porque así un tipo de usuario no podría visualizar ningún documento, si es que éste no tiene permiso para

[8] Por ejemplo, la Dirección General de Bibliotecas de la Universidad Nacional Autónoma de México (DGB) permite dentro de la UNAM ver tesis completas en formato electrónico, Si un usuario intenta visualizar una tesis fuera de la Universidad, es necesario proporcionar un nombre de usuario y contraseña.

hacerlo. Es decir, limitamos de alguna forma el uso del CHEM con el objetivo de conservar la seguridad de los documentos.

Entre los permisos de documento que se considerarán para esta tesis podemos mencionar los siguientes: documentos con permiso, documentos sin permiso, documentos con permiso en trámite, documentos que no requieren permiso y documentos con permisos parciales. Más adelante, en el capítulo correspondiente, hablaremos sobre estos permisos y su relación con los tipos de usuario. Por el momento, para ejemplificar la asociación que debe existir entre el tipo de usuario y el permiso de los documentos ponemos como ejemplo que un usuario administrador podría visualizar todos los documentos con cualquier permiso, mientras que un usuario de tipo básico únicamente tendría acceso a visualizar los documentos de los que se tenga permiso.

Siguiendo con la seguridad de los documentos, sabemos que Internet se ha convertido en una forma de satisfacer la necesidad de conocimiento; sin embargo, diariamente surgen nuevas formas de quebrantar las restricciones y las medidas que protegen a la misma volviéndola vulnerable. Por consiguiente, es fundamental desarrollar y aplicar métodos adecuados de seguridad a los documentos que se van a presentar, precisamente en este punto es donde tiene cabida la segunda forma de protección a los documentos, la de restringir y negar la posibilidad de copia e impresión de los mismos. Para esto se investigarán formas para protección de documentos haciendo uso de algún software.

En otro tema, los usuarios del corpus, ya sean investigadores o académicos, requieren de varias herramientas para analizar un corpus, no sólo de un generador de concordancias. Por ello, esta tesis contempla la adhesión de una herramienta más. Nos referimos a un generador de estadísticas, el cual obtendrá medidas cuantitativas de las concordancias haciendo uso del corpus. Las medidas que se tienen contempladas para presentar a los usuarios son las siguientes:

- Información mutua (I)
- Razón de semejanza ($-2\log\lambda$)
- Prueba de independencia (X^2)

- Coeficiente de coligación de Yule (Y)

Este generador de estadísticas ayudará en gran medida a entender la relación que existe entre la palabra de búsqueda y las palabras que la rodean ya sea antes, después o en ambos lados. Parte de esta tesis será el investigar dichas medidas y su aplicación en el procesamiento de corpus lingüísticos.

Regresando a la descripción de la primera versión del CHEM, una vez que el usuario accede a la página utilizando su nombre de usuario y contraseña, su acceso es registrado en un "libro de visitas". Éste refleja también el siguiente problema: el libro de visitas registra todos los accesos al sistema sin importar que el usuario realmente haga uso o no de la herramienta de explotación del corpus.

Del párrafo anterior surge la necesidad de perfeccionar el libro de visitas, para convertirlo en una herramienta capaz de administrar todas aquellas ocasiones en que el CHEM es utilizado tanto por usuarios registrados como por usuarios no registrados. El nuevo libro de visitas también deberá proporcionar información confiable y útil al equipo de desarrollo, específicamente al administrador del CHEM. Lo anterior con la finalidad de conocer con cierta precisión el alcance y uso del corpus.

Por último, quisiéramos referirnos a un tema que en la primera versión parece no ser un problema pero que sin duda lo será en la nueva versión: la velocidad. Al finalizar esta tesis el CHEM tendrá nuevas herramientas y opciones para su administración, por lo que será necesario agilizar de alguna manera la presentación de los mismos. Con el fin de disminuir, en lo posible, el tiempo de respuesta, esta investigación propone la utilización de un conjunto de tecnologías denominadas AJAX que posteriormente serán descritas.

Lo anterior refleja básicamente el funcionamiento actual del CHEM y deja ver que aunque de cierto modo se ha venido cumpliendo con su objetivo principal, aún falta mucho por alcanzar. Por ello, esta tesis busca aportar nuevas herramientas y con ellas una forma para administrarlas, aunque por supuesto el corpus requerirá después de más herramientas, contenidos y mantenimiento. Sólo estamos haciendo, como bien lo dice el título de la misma, un aporte tecnológico al Corpus Histórico del Español en México; aún con esta investigación quedará mucho por hacer.

0.2 Objetivo General

El objetivo general de esta tesis consiste en demostrar que la adecuada elección y aplicación de herramientas y métodos tecnológicos ayudarán al CHEM a cumplir con el objetivo por el que fue creado. Al mismo tiempo, esto solucionará los problemas que surgieron de su primera versión.

0.3 Objetivos específicos de investigación

Del objetivo general se desprenden los siguientes objetivos específicos de investigación:

- Investigar acerca de las fórmulas estadísticas para análisis de corpus: información mutua, razón de semejanza, prueba de independencia y coeficiente de coligación de Yule.
- Investigar métodos que nos permitan desarrollar el proceso para la obtención de las medidas estadísticas.
- Investigar métodos y herramientas para envío de correo electrónico para el registro de los usuarios.
- Investigar métodos para reconocer la dirección IP de los usuarios del CHEM.
- Investigar métodos y herramientas para permitir mostrar documentos completos en Internet.
- Investigar métodos para protección de documentos en Internet.
- Investigar métodos para validar la información que se proporciona al CHEM tanto para comprobar su fidelidad como para que ésta no represente una amenaza al corpus.

0.4 Objetivos específicos prácticos

Los objetivos específicos prácticos que contempla esta tesis se listan a continuación:

- Desarrollar una nueva herramienta de exploración del CHEM, que permita la generación de estadísticas mediante diversas fórmulas: información mutua, razón de semejanza, prueba de independencia y coeficiente de coligación de Yule.
- Utilizar una herramienta para envió de correo electrónico en la opción de registro de usuarios, que proporcione mayor veracidad de la información de los mismos.
- Tipificar a los diferentes usuarios que interactúan con el CHEM, permitiéndoles el uso o no de las herramientas del corpus.

- Proporcionar opciones de administración de los usuarios del CHEM considerando los tipos previamente definidos. Además de considerar la adición futura de nuevos tipos de usuarios.

- Reconocer a aquellos usuarios que utilicen el CHEM dentro de la UNAM, permitiéndoles el uso restringido de herramientas de análisis.

- A partir de la investigación realizada, validar la información que es proporcionada por los usuarios del CHEM.

- Conseguir que la aplicación muestre los documentos completos a determinados tipos de usuarios.

- Aplicar restricciones de seguridad para la protección de los documentos que permitirá visualizar la aplicación.

- Proporcionar información útil y confiable acerca del alcance y uso del CHEM optimizando la herramienta "libro de visitas" creada para la primera versión.

- Aplicar AJAX al CHEM para proporcionarle una mayor capacidad de respuesta y presentación.

0.5 Hipótesis

Este trabajo de tesis tiene como principal hipótesis de investigación la siguiente:

- El desarrollo de aplicaciones informáticas puede ayudar a facilitar la administración y el análisis estadístico de corpus lingüísticos electrónicos.

Del enunciado anterior se desprenden además, una serie de hipótesis alternas que también se busca comprobar. Las hipótesis se listan a continuación:

1. El uso de las estadísticas: información mutua, razón de semejanza, prueba de independencia y coeficiente de coligación de Yule permitirá que el CHEM muestre información sobre la relación entre palabras dentro de las concordancias agregando valor a las investigaciones que se puedan realizar.

2. La identificación de tipos de usuarios brindará una mejor administración de la aplicación del CHEM al permitir asignar diferentes herramientas a diferentes usuarios.

3. El uso del envío de correo electrónico, para confirmar el registro de los usuarios, permitirá que exista una mayor seguridad en cuanto a que el registro es adecuado y es hecho por una persona.

4. Tecnologías como Java Script así como algunos métodos de seguridad (por ejemplo, los que impiden las inyecciones SQL) pueden ser de gran ayuda para validar que la

información que es proporcionada a la aplicación sea en buena medida completa, correcta y veraz.

5. Existe algún método eficaz para identificar la dirección electrónica de los usuarios que utilicen el CHEM dentro o fuera de la UNAM.

6. El mostrar por completo los documentos del CHEM proveerá de mayor Información lingüística a las investigaciones que sobre el corpus puedan realizarse.

7. Las mejoras a la herramienta "libro de visitas" ayudará a que éste proporcione información que para el administrador del CHEM reflejará el alcance y uso que esté teniendo la aplicación.

8. La aplicación de un conjunto de nuevas tecnologías como AJAX, proporcionará al CHEM una mayor capacidad de respuesta en cuanto a tiempo y presentación.

0.6 Alcance de la Investigación

En lo que se refiere al alcance de esta tesis, como se ha mencionado ya en los objetivos, se proveerá al CHEM de una nueva herramientas de análisis de corpus lingüísticos: un generador de estadísticas. También, se optimizará el libro de visitas y se permitirá a algunos usuarios ver los documentos electrónicos en su totalidad. Además, al concluir esta tesis, el CHEM brindará nuevas posibilidades para administrar a sus usuarios y será mostrada una nueva interfaz que además será optimizada utilizando AJAX.

Esta tesis no se involucrará en ningún momento con la forma en que son obtenidas las concordancias por medio del generador de concordancias. El desarrollo de esta herramienta está a cargo de un grupo de becarios y desarrolladores del GIL y únicamente será referido en esta tesis como una herramienta ya realizada.

En lo que se refiere a los documentos del CHEM, esta tesis no planea escanearlos, transcribirlos o etiquetarlos, estas tareas están igualmente comisionadas a un grupo de becarios. Por lo anterior, los documentos serán considerados para esta investigación únicamente cuando las actividades mencionadas hayan sido completadas.

El generador de estadísticas solamente mostrará los resultados de las medidas aplicadas a las concordancias, las conclusiones que puedan surgir de los mismos dependerán del tratamiento que de ellas haga el investigador. En esta tesis, no se presentarán conclusiones de tipo lingüístico que pudieran surgir de estas estadísticas.

En cuanto a la tecnología utilizada para el desarrollo de esta tesis, resaltamos que la misma había sido previamente definida por el GIL, por lo que no se intentó evaluar un cambio, únicamente, como se mencionó en los objetivos, se planea integrar AJAX a la aplicación.

Con el fin de ser más específicos, listamos los recursos tecnológicos con los que cuenta el Grupo de Ingeniería Lingüística y que usará esta investigación, además, agregamos para cada uno de ellos una breve descripción:

- Java: es un lenguaje de programación orientado a objetos, desarrollado en Sun Microsystems por James Gosling. Java fue diseñado para distribuir contenidos a través de una red y su principal característica es que puede operar independientemente de la plataforma y del sistema operativo gracias al JVM (Java Virtual Machine).

- Java Server Pages (JSP's): un JSP es una tecnología de Java que permite generar contenido dinámico para web. Permite utilizar código JAVA mediante scripts. El código de un JSP es compilado como cualquier clase de Java. El rendimiento de una página JSP es el mismo que tendría el servidor equivalente, ya que el código es compilado como cualquier otra clase Java. A su vez, la máquina virtual compilará dinámicamente a código de máquina las partes de la aplicación que lo requieran. Esto hace que JSP tenga un buen desempeño y sea más eficiente que otras tecnologías web que ejecutan el código de una manera puramente interpretada.

- Servlets: Un servlet es un pequeño programa que se ejecutan del lado del servidor. Los servlets extienden dinámicamente la funcionalidad de un servidor Web.

- Tomcat: es un software de código abierto que sirve como contenedor de servlets con soporte para interpretar JSP's

- PostgreSQL: es un sistema manejador de bases de datos objeto-relacional.

- DHTML: estas siglas significan Dynamic HTML o HTML dinámico, es una combinación de HTML, hojas de estilo (Cascading Style Sheets, CSS por sus siglas en inglés) y Java Script que brinda la posibilidad de crear sitios web interactivos.

- Java Script: es un lenguaje de programación que al poder ser embebido en páginas WEB permite la creación de sitios interactivos. Este lenguaje es interpretado y ejecutado en el navegador del cliente al ser cargada una página.

- AJAX: AJAX significa Asynchronous Java Script + XML y es la unión de varias tecnologías para lograr la comunicación asíncrona entre el servidor de aplicaciones y el cliente.

0.7 Metodología

Esta tesis comprende un tipo de investigación aplicada, debido a que tiene como finalidad la consolidación y aplicación de conocimientos generados para las áreas de informática, lingüística de corpus, computación y estadística.

Es también una investigación descriptiva y documental, la primera porque busca describir las características destacadas de cada parte de la investigación para poner de manifiesto su estructura o comportamiento; y documental porque ha sido realizada partiendo de fuentes como libros, revistas y publicaciones en línea para permitir el análisis y estudio necesarios.

Para aclarar la metodología indicamos a continuación los pasos que se decidieron llevar a cabo, todos ellos dentro de alguna de cuatro categorías que proponemos para facilitar su comprensión.

1. *De Investigación y redacción*

- Elaborar el marco teórico-conceptual de la tesis con el fin de conocer los conceptos fundamentales de la lingüística de corpus.
- Conocer a fondo el proyecto del CHEM.
- Identificar los tipos de usuarios que reconocerá el CHEM.
- Identificar métodos para validar el formato de la información que es capturada.
- Investigar diferentes opciones para presentar el Acuerdo General de Usuarios del CHEM.
- Investigar herramientas o métodos para el envío de correo electrónico.
- Investigar métodos para generar cadenas de números aleatorios en Java.
- Investigar métodos y herramientas para la identificación de la IP del cliente.
- Investigar métodos para presentar menús desplegables con Java Script y CSS.
- Identificar, por medio de la dirección IP, si un usuario utiliza el CHEM dentro de la UNAM.
- Identificar los tipos de permisos de los documentos.

- Investigar métodos y herramientas que permitan mostrar documentos electrónicos en internet.
- Investigar métodos para proteger los documentos que mostrará la aplicación.
- Investigar acerca de las medidas estadísticas que podrán realizarse en el CHEM (información mutua, razón de semejanza, prueba de independencia y coeficiente de coligación de Yule).
- Investigar métodos para programar las medidas estadísticas.
- Identificación de casos que pueden ser registrados en el libro de visitas.

2. *De Implementación*

- Diseño de la base de datos.
- Diseño de la nueva interfaz del CHEM.
- Implementar métodos para validar el formato de la información.
- Elegir e implementar opción para presentar el Acuerdo General a los usuarios.
- Elegir e implementar el método para envío de correo electrónico.
- Implementar método de generación de número aleatorio.
- Implementar método para permitir a los usuarios recuperar su contraseña mediante el envío de un correo electrónico.
- Implementar el método para la identificación de la IP del cliente.
- Elegir e implementar método para presentar menús desplegables con Java Script y CSS.
- Elegir e implementar método para mostrar documentos electrónicos del CHEM por medio de la aplicación en internet.
- Implementar métodos para proteger los documentos que mostrará la aplicación.
- Implementar los métodos para generar las medidas estadísticas.
- Diseño de la interfaz del Administrador del CHEM.
- Diseño de base de datos para la gestión de los usuarios y herramientas del corpus.

- Implementación de los métodos para permitir al Administrador la gestión de la aplicación.
- Diseño de base de datos para registrar accesos en el libro de visitas.
- Implementación del libro de visitas para usuarios registrados.
- Implementación para libro de visitas para usuarios no registrados.

3. *De pruebas y corrección*

- Probar y corregir métodos para validar el formato de la información.
- Probar y corregir método de envió de correo electrónico.
- Probar y corregir método de generación de número aleatorio.
- Pruebas y corrección al registro de usuario.
- Probar método para la identificación de la IP del cliente.
- Probar y corregir menú desplegables con Java Script y CSS.
- Probar y corregir método para mostrar documentos electrónicos del CHEM por medio de la aplicación en internet.
- Probar y corregir métodos para proteger los documentos que mostrará la aplicación.
- Probar y corregir los métodos para generar las medidas estadísticas.
- Pruebas y corrección de los métodos para permitir al Administrador la gestión de la Aplicación.
- Pruebas y corrección para libro de visitas de usuarios registrados.
- Pruebas y corrección para libro de visitas de usuarios no registrados.

4. *De integración*

- Integrar métodos de validación de información al registro del usuario.
- Integrar método de generación de número aleatorio al correo electrónico.

- Integrar el envío de correo electrónico al registro de usuario para que concluya el proceso.
- Integrar el menú desplegable a la nueva interfaz del CHEM.
- Integrar el proceso de registro de usuarios del CHEM a la nueva interfaz.
- Integrar el módulo que permitirá mostrar los documentos completos a la aplicación.
- Integrar el generador de estadísticas al CHEM.
- Integrar el libro de visitas al CHEM.

1. La Ingeniería Lingüística

El presente capítulo se centra en definir los conceptos de lingüística e ingeniería lingüística, así como en describir los diferentes niveles del lenguaje, útiles para desarrollar modelos para el análisis lingüístico. También agregamos la descripción de algunas áreas del procesamiento del lenguaje natural.

1.1 Lingüística y Lingüística Computacional

Comenzamos definiendo a la lingüística como la ciencia que estudia el lenguaje humano y como cualquier otra ciencia, ésta construye los modelos y las descripciones de su objeto de estudio: el lenguaje natural. Por su parte, la lingüística computacional trata de la construcción de modelos del lenguaje "entendibles" para las computadoras **(Gelbukh y Sidorov, 2006: 5)**. La ciencia lingüística se divide en varios niveles de análisis que veremos a continuación.

1.1.1 Niveles de análisis lingüístico

Para su estudio, la lingüística se divide en varios niveles de análisis, en lingüística computacional, estos niveles del leguaje ayudan a crear modelos lingüísticos que una computadora pueda ser capaz de procesar. Estos niveles además, no son más que partes de un modelo completo del lenguaje. Tradicionalmente, el lenguaje puede dividirse en seis niveles, mismos que listamos a continuación y que posteriormente describimos:

1. Fonética/fonología.
2. Morfología.
3. Sintaxis.
4. Semántica.
5. Pragmática.
6. Discurso.

- **Fonética/fonología.**

La fonética es la parte de la lingüística encargada de la exploración de las características de los sonidos del lenguaje, por su parte la fonología estudia los fenómenos basados en unidades

lingüísticas abstractas llamadas fonemas. Ambas también estudian los sonidos de los diferentes idiomas, sus relaciones y sus implicaciones.

Para ejemplificar la aplicación de la fonética y de la fonología podemos referirnos a los sistemas de reconocimiento de voz que existen hoy en día. Estos sistemas hacen posible, por medio de una computadora, identificar las palabras pronunciadas por un humano; sin embargo, existen todavía muy pocos capaces de reconocer todas las palabras de diferentes tipos de hablantes.

Otro ejemplo son los programas de síntesis de habla, que son aquellos que pueden ser capaces de convertir texto en habla. Cabe desatacar que este tipo de programas son más exitosos en cuanto a su funcionalidad, aunque su área de aplicación es más restringida debido a que suelen ser más útiles para personas con alguna discapacidad visual.

- **Morfología**

La morfología se encarga de estudiar la estructura interna de las palabras, por ejemplo, la forma en que se unen sufijos, prefijos y raíces para formar palabras. Además, comprende la manera como se expresan los sistemas de categorías gramaticales, como género y número, en los diferentes idiomas.

En la lingüística computacional resulta especialmente complicado desarrollar sistemas de análisis morfológico automático debido a que, por ejemplo, se requieren de diccionarios con grandes volúmenes de raíces. Derivado de lo anterior, aunque existen algunos sistemas funcionales para muchos idiomas, no ha sido posible identificar un estándar para el análisis morfológico.

- **Sintaxis**

La sintaxis atiende al análisis de las relaciones que existen entre las palabras dentro de la frase, donde dichas relaciones se pueden representar como dependencias o constituyentes **(Gelbukh y Sidorov, 2006: 66)**.

Entonces, si quisiéramos por ejemplo construir un sistema computacional para el estudio de la sintaxis, el sistema debería por lo tanto tener métodos de análisis y síntesis

automática para que sea capaz de construir la estructura de una frase (análisis) o generar una frase basándose en su estructura (síntesis).

- **Semántica**

Este nivel se encarga de entender una frase, es decir, conocer el sentido de las palabras que la componen junto con sus relaciones sintácticas. La semántica incluye disciplinas como: lexicología y lexicografía, las cuales definen los sentidos de las palabras. Resulta interesante que muchas veces, el resultado de esa definición es un círculo vicioso debido a que las palabras se definen a través de otras palabras.

Una de las intenciones de la semántica computacional es ayudar a resolver el problema del círculo vicioso, identificando un conjunto de palabras llamado *vocabulario definidor* para definir a todas las demás.

- **Pragmática**

La pragmática trata de las relaciones entre la oración y el mundo externo, se puede decir que lo que le interesa a la pragmática es la intención del autor del texto o del hablante.

- **Discurso**

El discurso es la unión y la relación de las oraciones que utilizamos cuando hablamos los seres humanos. Para poder analizar el lenguaje en este nivel, la correferencia es un concepto esencial. La correferencia se da cuando varias unidades del discurso hacen referencia a la misma entidad, por ejemplo, cuando utilizamos pronombres para sustituir un sustantivo o un sujeto. Por lo anterior, un sistema computacional que pretenda analizar el discurso debe tener la capacidad para interpretar la correferencia y construir representaciones semánticas a partir de ella.

Como hemos visto, los niveles del lenguaje son divisiones del mismo que nos sirven para hacer análisis y desarrollar modelos que puedan ser aplicados a los sistemas computacionales de análisis lingüístico **(Gelbukh y Sidorov, 2006: 63-69)**.

1.2 Ingeniería lingüística

Citando un documento publicado por la Comisión Europea[9], titulado *Ingeniería lingüística: Como aprovechar la fuerza del lenguaje* definimos a la ingeniería lingüística como: "la aplicación de los conocimientos sobre la lengua al desarrollo de sistemas informáticos que puedan reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas" **(WEB, 03)**.

La ingeniería lingüística trata entonces de los modelos computacionales del lenguaje natural y cómo hacer entender a las computadoras los idiomas humanos. Se basa además en el conocimiento del funcionamiento de la lengua para desarrollar soluciones que ayuden a entender y manipular el lenguaje, construyendo herramientas que permitan el análisis lingüístico y el desarrollo tecnológico **(WEB, 04)**. En los siguientes apartados exponemos algunas propiedades del procesamiento de lenguaje natural para ayudar a profundizar en el concepto de ingeniería lingüística**.**

1.2.1 Procesamiento de lenguaje natural (PLN) y lingüística computacional (LC).

El procesamiento del lenguaje natural se encarga de habilitar a las computadoras para entender textos y facilitar repuestas de los mismos, tratando de procesar un texto por su sentido y no sólo como un archivo **(Gelbukh y Sidorov, 2006: 15-16)**. Según estos autores, el esquema general para el procesamiento de lenguaje natural o lingüística computacional es el siguiente.

- El texto no se procesa directamente sino que se transforma en una representación con todas las características del original.
- El programa principal manipula esta representación, la transforma para buscar subestructuras necesarias.

[9] Para ser precisos el documento publicado por la Comisión Europea es un folleto que lleva por título Ingeniería lingüística. Cómo aprovechar la fuerza del lenguaje. Este folleto fue preparado por Anite Systems para el proyecto LINGLINK en nombre de los participantes del sector "Ingeniería Lingüística" del Programa de Aplicaciones Telemáticas y adaptado al español por el Observatorio Español de Industrias de la Lengua. Para obtener más información puede consultar la dirección especificada en la referencia WEB, 03.

- Por último, si se requiere, los cambios hechos a la representación formal se transforman en lenguaje natural.

El procesamiento del lenguaje natural se puede emplear en un amplio rango de tareas muy sencillas y rutinarias o en situaciones complejas. Según Gelbukh y Sidorov, en México desde los años setenta se ha reconocido la importancia del procesamiento del lenguaje natural en áreas como la ciencia, la educación, el comercio, la cultura, el gobierno y otros aspectos de la vida social como un factor crítico para la independencia cultural, técnica y económica de nuestro país.

Por esos años también, Luis Fernando Lara, investigador de El Colegio de México, comenzó a realizar investigaciones sobre técnicas estadísticas aplicadas al análisis automático de un corpus de español mexicano, con el fin de desarrollar un diccionario de las palabras usadas en México.

Por el año de 1996, en el Instituto Politécnico Nacional (IPN), se fundó el Laboratorio de Lenguaje Natural y Procesamiento de Texto el cual se ha encargado de desarrollar varios proyectos científicos y tecnológicos en las áreas de análisis sintáctico y semántico, aprendizaje automático de los recursos léxicos y compilación de diccionarios, minería de textos y resolución de anáfora.

En 1998, en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), de la UNAM, se formó el grupo de investigación en lingüística computacional. Éste grupo hace investigación en diálogos multimodales, así como en formalismos gramaticales modernos. De igual forma en la UNAM también se creó el GIL el cual se enfoca a resolver problemas lingüísticos desde un punto de vista más práctico.

1.2.2 Áreas del PLN y de la Ingeniería lingüística

La información se ha vuelto un recurso indispensable para los seres humanos, es fuente de conocimiento y su manejo adecuado permite tomar ventajas en diversos ámbitos sociales y culturales. Por su parte, el lenguaje ha sido uno de los principales medios para transmitir información y con ella el conocimiento, por ejemplo, en forma de documentos, libros, artículos, etc., que en la actualidad pueden ser conservados en formato digital. Sin embargo, una computadora no puede entender el significado de un texto, por lo que muchos países se

han esforzado en desarrollar ciencia que permita habilitar a las computadoras para entenderlo. En los siguientes apartados describimos algunas áreas del PLN por las que es importante desarrollar ciencia que permita el análisis lingüístico automático.

1.2.2.1Recuperación de Información

Los sistemas de recuperación de información (IRS, por sus siglas en inglés) son diseñados para buscar información particular en una serie de documentos contenidos en bases de datos o bancos textuales. Los primeros IRS fueron desarrollados para ser utilizados en la búsqueda de artículos científicos. Estos sistemas pueden obtener información utilizando diferentes métodos de búsqueda **(Bolshakou y Gelbukh, 2004: 63-66)**.

Usualmente, los científicos sustituyen los documentos por grupos de palabras clave, que resultan ser las palabras más importantes en un tema. Las palabras clave son adjuntadas al documento en la base de datos del IRS, por lo que una consulta puede tener como condición una o varias de esas palabras, de esta forma, el sistema recuperará los documentos donde se encuentre alguna de las palabras clave.

La forma en que un IRS realiza una búsqueda se puede describir como sigue:

- La consulta contiene un conjunto de palabras clave como condición.
- El sistema busca los documentos que contienen todas las palabras clave.
- El sistema busca los documentos que contienen todas las palabras clave excepto una y así sucesivamente, hasta que finalmente busca los documentos que contienen solo una de las palabras clave.
- Los resultados son ordenados y presentados por grado de relevancia. El grado de relevancia se refiere a números de palabras clave que fueron encontradas en un documento.

Según Bolshakou y Gelbukh, podemos hablar de dos conceptos importantes en un sistema IRS, estos son: el *recall* y la precisión. El primero es la relación del número de documentos relevantes encontrados dividido entre el número total de documentos relevantes

en la base de datos. Por su parte, la precisión es la relación del número de documentos relevantes divididos entre el número total de documentos encontrados.

Hace tres décadas, el problema de la extracción automática de palabras clave era llamado “abstracción automática”. Cabe destacar que las palabras más frecuentes en un texto de negocios, científico o técnico, son puramente auxiliares y no reflejan la verdadera esencia del texto. Sin embargo, actualmente un IRS es capaz de crear automáticamente un grupo de palabras clave a partir del texto de un documento completo (subsistemas de indexación).

Sin embargo, los resultados que proporcione el IRS de las operaciones de recuperación dependen directamente de la calidad y operación de los subsistemas de indexación y comparación, el contenido de la terminología del sistema y otros datos y capacidades del sistema.

1.2.2.2Extracción de información

Otra de las áreas del PLN es la que brinda la posibilidad de crear bases de datos a partir de grandes volúmenes de textos, estas bases pueden contener datos específicos acerca de la información que se comunica en los textos. Por ejemplo, una base de datos que guarde información de atracciones turísticas, lugares y servicios, a partir de la extracción de información de páginas Web y propagandas. Para esto, es importante que el sistema que va a crear una base de datos extrayendo información, de una o varias fuentes, tenga cierto grado de comprensión del texto sobre el que se haga el procesamiento.

En la extracción de información influyen dos conceptos: el filtrado y la alerta. El primero se refiere a seleccionar, de información nueva, sólo la que corresponde al interés del usuario, a través de un agente de filtrado. El segundo es útil cuando la información aparece con poca frecuencia. Lo que hace la alerta es advertir al usuario si llega a aparecer información de su interés[10].

[10] El ejemplo de un sistema de extracción de información que obtiene lanzamientos de software a partir de sitios Web se puede encontrar en la tesis: *Sistemas de extracción de información. Soluciones informáticas organizacionales basadas en datos no estructurados (2006)*.

1.2.2.3Traducción automática

La traducción automática permite romper barreras del lenguaje y facilita la comunicación entre personas de diferentes países y culturas. Ofrece también un acceso fácil a la información escrita contenida en libros y revistas de otras lenguas.

Esta área del procesamiento del lenguaje natural fue concebida en un principio con la idea de que una computadora pudiera sustituir las palabras escritas en un idioma a las palabras de otro; sin embargo, el resultado no era nada coherente ni legible. La lingüística computacional comienza sus estudios sobre la traducción precisamente con el análisis de la causa de este fenómeno.

La traducción automática presenta importantes retos para generar resultados correctos, por ejemplo, el orden de las palabras, que es muchas veces diferente en dos idiomas distintos; la ambigüedad que puede tener una frase; y o la selección de palabras e incluso el uso de algunas expresiones. Es necesario que un sistema de traducción automática haga uso adecuado de métodos para el análisis de texto y la representación de su contenido que le permitan comprenderlo para convertirlo correctamente con el sentido original **(Gelbukh y Sidorov, 2006: 39-42)**.

La forma correcta de traducir un texto es entenderlo como lo hace una persona. La forma en que una máquina puede traducirlo es a través del análisis lingüístico para después generar un texto con el sentido original del otro idioma. A continuación presentamos el esquema general que siguen la mayoría de los traductores según Gelbukh y Sidorov:

- El texto original se transforma en una representación intermedia.
- Si son necesarios, se pueden hacer algunos cambios.
- Por último, la representación intermedia se transforma al texto en el lenguaje final.

2. Lingüística de Corpus

En los siguientes apartados, enunciamos diferentes conceptos para poder entender en qué consiste la lingüística de corpus, definiendo para ello los conceptos de: corpus y corpus lingüístico, además señalando sus objetivos y características, así como su estrecha relación con la informática. Lo anterior con la finalidad de sentar la bases para una mejor comprensión del CHEM.

2.1 Lingüística de corpus

La lingüística de corpus es una parte de la lingüística que tiene por objetivo permitir el estudio de la lengua utilizando medios automatizados, los cuales permiten procesar grandes volúmenes de datos lingüísticos **(WB, 05)**. Para lograr su cometido, la lingüística de corpus hace uso de corpus lingüísticos.

2.2 Corpus

En estricto sentido etimológico, un corpus está definido como una recopilación de textos. La Real Academia Española añade además: "que debe ser una colección lo más extensa y ordenada posible de textos científicos, literarios, etc., que puede servir de base a una investigación" **(WB, 06)**. Otra definición nos dice que un corpus es una muestra de la lengua ya sea oral, escrita, etc. **(Blecua et al, 1999: 52)**.

Es importante mencionar que un corpus no está únicamente compuesto por las reflexiones de teóricos o por materiales cuidadosamente redactados, también puede integrarse por una gran variedad de materiales como: conversaciones diarias, publicaciones escritas, publicaciones infantiles, cartas, programas de radio, etc. Todos estos datos se agrupan en textos que pueden ser utilizados para múltiples propósitos de investigación en un gran número de disciplinas es decir, los corpus son un recurso multifuncional.

2.3 Corpus lingüísticos

Según Tony McEnery, un reconocido lingüista inglés, un corpus es una gran cantidad de evidencia lingüística típicamente compuesta de ejemplos del uso del lenguaje **(2003: 449).** Por su parte para J. Sinclair, un especialista en el campo de los corpus modernos, un corpus es: una colección de segmentos de la lengua que son recopilados y ordenados de acuerdo a

criterios lingüísticos explícitos con el objetivo de ser utilizados como una muestra del lenguaje **(1994:4).**

Detallamos, que un corpus lingüístico además de ser un conjunto de recopilaciones textuales debidamente ordenadas, debe cumplir al menos con dos características: representatividad y equilibrio. La primera característica se refiere a que un corpus sólo puede contener una muestra representativa de las variaciones de la lengua que tiene la población de estudio. El equilibrio por su parte, se relaciona con la proporcionalidad del contenido de un corpus el cual, debe evitar ser tendencioso hacia solo ciertos materiales.

El objetivo principal de un corpus lingüístico es permitir el análisis lingüístico, porque su importancia reside en la posibilidad que ofrece para manipular grandes cantidades de datos recopilados de una amplia diversidad de hablantes, es decir, son un medio de procesamiento de datos lingüísticos.

Para resumir, señalamos que un corpus lingüístico no es cualquier colección de datos lingüísticos que se componga de tres oraciones o de miles de palabras. Un corpus lingüístico es, como se mencionó, una colección de datos bien organizados bajo ciertos criterios y que además cuenta con una interfaz para permitir la exploración de ciertas características o grupo de características lingüísticas **(McEnery, 2003: 449)**.

2.3.1 Tipos de corpus

En la realidad es prácticamente imposible que un solo corpus pueda contener todas las expresiones del lenguaje natural. Es necesario definir algunos tipos de corpus con la finalidad de limitar su estructura de acuerdo a la procedencia de su contenido y a los objetivos que se busquen cumplir con su realización. Entre los tipos de corpus que pueden existir distinguimos a los siguientes:

- **Corpus hablados**

Este tipo de corpus está enteramente compuesto por lenguaje hablado, es decir, en principio se compone de un conjunto de grabaciones que, por otra parte, pueden ser transcritas para estar disponibles como la única fuente de datos para analizar, dejando de lado las

grabaciones. Sin embargo, el uso de la transcripción o de la grabación tiene inconvenientes, ya que si el corpus existe sólo como grabaciones de sonido, el análisis de los datos se complica.

Por ejemplo, cuando queramos buscar la ocurrencia de la palabra *manzana*, es problemático para la máquina hacer búsquedas en un corpus hablando donde están representados una gran variedad de hablantes, ya que los rasgos acústicos de cada hablante serán diferentes. Por otra parte, si únicamente se encuentra disponible la transcripción del corpus hablado, se perderían muchas características acústicas importantes. Una solución para evitar tener problemas con un corpus hablado, es que éste debe ayudarse de la transcripción del mismo para obtener mejores resultados a la hora de analizar los datos.

- **Corpus monolingüe y multilingüe**

El corpus monolingüe representa un sólo idioma, mientras que el multilingüe se compone de una variedad de lenguas.

- **Corpus comparable**

Es un conjunto de corpus monolingües de diversos idiomas, cada uno con ciertas características o grupo de características lingüísticas similares. Este tipo de corpus permite, dada su estructura, el estudio de datos lingüísticos con sus diferentes contrastes.

- **Corpus paralelo**

Este corpus se compone primero por textos en un mismo idioma que después son traducidos a uno o más idiomas. Por lo general, cada línea de un corpus tiene una correspondencia con una línea del otro.

- **Corpus sincrónico y diacrónico**

Una definición para este tipo de corpus, es la que encontramos en el curso *Lingüística de Corpus* de Sierra **(WB, 07).** Éste menciona que los corpus sincrónicos contienen textos de un momento específico en el tiempo, mientras que los corpus diacrónicos son los que comparan, confrontan o relacionan muestras lingüísticas a través de varios periodos de tiempo.

Un ejemplo de ello son los dos corpus de la Real Academia Española. El primer, uno sincrónico que contiene documentos de 1975 hasta la actualidad, y el segundo, uno diacrónico que se compone de documentos elaborados antes de 1975. Más adelante retomaremos el tema de los corpus lingüísticos construidos por la Real Academia Española.

- **Corpus textual, de referencia y léxico**

El corpus textual incorpora íntegramente los documentos que lo componen, el corpus de referencia únicamente recopila fragmentos de documentos, mientras que el corpus léxico reúne fragmentos de tamaño constante para poder analizar el vocabulario del corpus.

- **Corpus canónico**

Está formado por todos los textos que conforman la obra de un autor, independientemente de los géneros.

- **Corpus General**

Este tipo de corpus intenta reflejar la lengua común de una colectividad en un ámbito más amplio, porque no le interesa recoger materiales de un sólo género.

- **Corpus Especializado**

A diferencia del corpus general, este corpus recoge textos que pueden aportar datos para la descripción de un tipo particular de lengua.

2.4 Corpus lingüísticos electrónicos

Los corpus electrónicos han existido desde finales de 1940. Sin embargo, aunque por esos años ya se tenía la idea de estudiar el lenguaje utilizando recopilaciones de textos, también existía el problema de la limitada capacidad que tenían las computadoras para el almacenamiento y manipulación de grandes volúmenes de datos; esto volvía una tarea casi imposible el poder procesar de alguna forma todos esos datos lingüísticos.

Los corpus lingüísticos electrónicos, que hacen uso de técnicas interactivas y automáticas basadas en el uso de las computadoras para analizar con relativa facilidad datos lingüísticos, son en realidad un fenómeno reciente.

2.4.1 Definición

Los corpus lingüísticos electrónicos son un conjunto de materiales escritos o hablados bien organizados para permitir el estudio y exploración de ciertas características de la lengua, sin embargo, la diferencia entre estos y un corpus lingüístico radica en que el primero hace uso de las herramientas que la tecnología le ofrece para poder procesar y analizar los datos, dando precisión, rapidez y confiabilidad a la información que resulta.

Es importante mencionar que el uso de computadoras para realizar análisis lingüísticos sobre corpus aporta muchas ventajas para el análisis del lenguaje. Algunas de ellas son:

- La posibilidad de identificar y analizar complejos patrones del uso del lenguaje.
- El permitir el almacenamiento y análisis de grandes bases de datos de leguaje natural.
- La consistencia y fiabilidad de los análisis, agregadas al procesamiento de los datos.

Otra serie de ventajas adicionales, pero no menos importantes, según Sierra son **(WEB, 08)**:

- La facilidad para la manipulación del contenido del corpus.
- La rapidez para la obtención de información resultado del procesamiento del corpus.
- La precisión en la información que genera.

Un corpus lingüístico electrónico por otra parte también presenta un conjunto de desventajas, entre las cuales destacan:

- La digitalización de los textos puede ser una tarea que consume demasiado tiempo.
- Las características del equipo que contenga y analice el corpus deben ser suficientes para contenerlo y procesarlo.
- Si ocurre alguna falla técnica, el corpus puede tener problemas de disponibilidad o vulnerabilidad de la información.

En cuanto a la evolución que ha tenido el desarrollo de corpus lingüísticos, podemos mencionar, que en su historia ha habido importantes hitos que han permitido el surgimiento de nuevos corpus y el aumento de tamaño de los ya existentes, así como el desarrollo de herramientas computacionales para su análisis. Entre los hitos, podemos señalar los siguientes:

En 1964, el desarrollo del Corpus de la Universidad de Brown (*Brown University Corpus*), el primer corpus en formato electrónico, y la articulación de conceptos claramente relacionados con las ideas de equilibrio y representatividad de los corpus actuales.

En 1980, la aparición de la transcripción por Svartvik y Quirk, ambos estudiosos de la lingüística, de un corpus hablado.

En 1990, la creación del Corpus Nacional Británico (*British National Corpus*) y el Banco de la lengua inglesa (*Bank of English*) que contenía ya 300, 000,000 palabras del inglés moderno, una tarea prácticamente imposible para las primeras computadoras. En ese mismo año, el incrementó del carácter multilingüe de la lingüística de corpus aunado al aumento de la disponibilidad de corpus monolingües en diversos idiomas y el uso generalizado de corpus paralelos **(McEnery, 2003: 452).**

Por otro lado, desde la invención de los programas de reconocimiento óptico de caracteres (OCR, por sus siglas en inglés), la tendencia ha sido digitalizar los textos para poder ser procesados electrónicamente. Hoy en día este tipo de herramientas ha ayudado a extender el uso de los corpus lingüísticos electrónicos para ayudar no sólo a facilitar análisis lingüísticos, sino también a desarrollar tecnologías del lenguaje como traductores, diccionarios, bases de datos, entre otros. Estas tecnologías no serían tan eficientes sin haber recurrido al uso de corpus.

2.4.2 Corpus existentes

El uso de un corpus constituye una herramienta para facilitar estudios lingüísticos. Es por eso que durante años se han venido desarrollando en muchas partes del mundo diferentes corpus, hasta que en la actualidad es común encontrar corpus orientados a diversos tipos de

análisis. A continuación hacemos referencia a algunos de los Corpus más reconocidos, ya sea por su extensión, diseño u objetivos.

2.4.2.1Corpus de Referencia del Español Actual (CREA) y Corpus Diacrónico del Español (CORDE)

Entre los corpus lingüísticos más populares, tanto por su contenido como por su tamaño, se encuentran los dos desarrollados por la Real Academia Española que a continuación exponemos.

CREA

El Corpus de Referencia del Español Actual (CREA) está siendo construido por la Real Academia Española desde 1996, es un corpus monolingüe compuesto en un 90% por una amplia variedad de textos escritos completos y un 10% de trascripciones orales producidos en todos los países de habla hispana. Para el año 2004 el CREA contenía 170 millones de registros.

En cuanto a su constitución, el 50% del volumen del CREA está compuesto por material perteneciente a España y el otro 50% por materiales de procedencia hispanoamericana divididos en seis zonas: andina, caribeña, central, chilena, mexicana y rioplatense. Contiene además, recopilaciones integras de libros, prensa, revistas y transcripciones de documentos sonoros, en su mayoría provenientes de la radio y televisión **(WB, 09)**.

El proceso para introducir un texto en el CREA involucra una serie de pasos que puntualizamos de la siguiente forma:

Primero es escanear el documento, luego corregir posibles errores cometidos por el programa de reconocimiento óptico de caracteres (OCR) y posteriormente introducir marcas según el estándar de marcado SGML (*Standard General Markup Language*), que permite después la recuperación de la información[11].

[11] SGML (*Standard General Markup Language*), es un estándar anterior a lo que hoy es XML (Extensible Markup Language), recomendado por la TEI (*Text Encoding Initiative*). La TEI se encarga de producir lineamientos para el marcado de textos de uso académico de humanidades, dichos lineamientos de TEI

Luego, se utilizan macros desarrolladas en Word para la inserción de las marcas y en el caso de los materiales periodísticos, se agrega el área temática. Hasta el 2001, el CREA añadía dos tipos de marcas: estructurales (párrafo, oración, número de página) y de resalte tipográfico (negritas, cursivas, texto entrecomillado). También se llena la cabecera del texto con los datos bibliográficos.

Finalmente se valida como texto SGML y se genera una copia en formato texto, que se integra al corpus. Para el caso de introducir documentos transcritos de grabaciones al corpus oral, se utiliza XML **(WB, 10).**

CORDE

El Corpus Diacrónico del Español (CORDE) es un corpus con materiales únicamente en lengua española, elaborados desde el inicio del idioma hasta 1975. Comenzó a elaborarse en 1994, con la idea de aplicar técnicas informáticas para construir un banco de datos para la Real Academia Española. El objetivo del CORDE es servir a la extracción de información para estudiar las palabras y sus significados, así como la gramática y su uso a través del tiempo.

Hasta el año 2005, el CORDE contaba con 250 millones de registros, lo que lo convierte en el corpus de mayor contenido en la historia de la lengua española. Los materiales que lo construyen están distribuidos en prosa y verso y, dentro de cada modalidad, en textos narrativos, líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos, periodísticos, etc.

En cuanto al origen de los textos, estos provienen de libros escaneados, otros están en formato electrónico y algunos transcritos en formato digital ya que no existían ediciones modernas de ciertas obras. Al igual que con el CREA, para el CORDE es utilizado un programa de reconocimiento de caracteres (OCR) que genera el documento electrónico, después se agrega una serie de marcas textuales SGML para permitir la recuperación de la información y el intercambio de textos con otros corpus **(WB, 11)**.

---

proveen un formato estándar pero, eso no significa que un corpus deba ser marcado con todos los detalles proporcionados por ese formato.

2.4.2.2Corpus del español de Mark Davies

Otro de los corpus que se ha alcanzado gran reconocimiento es el Corpus del Español de Mark Davies, el cual contiene 100 millones de palabras del español de los siglos XIII al XX. A diferencia de otros corpus del español, el Corpus del Español permite realizar búsqueda por 35 categorías gramaticales, 20,000 lemas, y 30,000 grupos de sinónimos y antónimos, además de búsquedas por etimología, frecuencia, y por categorías semánticas y sintácticas creadas por el usuario mismo **(WB, 12).**

La arquitectura del Corpus del Español está compuesta por varias bases de datos relacionales en SQL Server y permite realizar búsquedas con una velocidad de menos de 2-3 segundos. Este corpus también intenta resolver algunos de los problemas que son parte del CORDE, como que se puedan realizar búsquedas complejas y usando comodines para palabras con cierta terminación.

2.4.2.3Corpus Lingüístico de Ingeniería (CLI) y Corpus de las Sexualidades en México (CSMX)

El GIL desarrolla, además del CHEM, dos corpus sincrónicos especializados, que hacen uso de las mismas técnicas y herramientas construidas por el grupo para su análisis. El primero, el Corpus Lingüístico en Ingeniería (CLI), incluye documentos de distintas áreas de la ingeniería como la ingeniería eléctrica, mecánica y civil. El segundo, el Corpus de las Sexualidades en México (CSMX), es un conjunto de documentos de distintas fuentes relacionados con el tema de la sexualidad. Ambos corpus serán utilizados para el trabajo de extracción automática de términos, entre otras actividades[12].

2.4.2.4Otros corpus

British National Corpus

El Corpus Nacional Británico (BNC, por sus siglas en inglés) es un corpus monolingüe y sincrónico que contiene 100 millones de palabras, muestras de la lengua escrita y hablada con

[12] De hecho, las aportaciones de esta tesis impactarán de forma directa al desarrollo de estos corpus.

el objetivo de representar una amplia sección del inglés británico de la última parte del siglo XX. El corpus ha sido codificado de acuerdo al estándar recomendado por la TEI.

El 90% del corpus está comprendido por la lengua escrita, derivada de periódicos regionales y nacionales, revistas especializadas para todas las edades e intereses, libros académicos y ficción popular, publicados y no publicados, cartas y memorandos, ensayos universitarios, entre otros. Mientras que el 10% lo conforma la lengua hablada, con transcripciones de conversaciones informales y recopilaciones del lenguaje hablado en diferentes contextos.

El desarrollo del BNC comprende los años de 1991 hasta 1994. Posteriormente no se agregaron más textos pero ha sido revisado ligeramente antes del lanzamiento de su segunda edición en el 2001 (BNC World) y para la tercera edición en el 2007 (BNC XML Edition) **(WB, 13).**

Corpus del IULA

El corpus del Instituto Universitario de Lingüística Aplicada (UILA) es un corpus textual y específico que recopila textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) recopilados por especialistas en las áreas de economía, derecho, medio ambiente, medicina e informática **(WB, 14).**

CUMBRE

El Corpus Lingüístico del Español Contemporáneo (CUMBRE) contiene 20 millones de palabras del español oral y escrito de España e Hispanoamérica. Se constituye de textos extraídos de libros diversos como: novela policiaca, novela histórica, política, deportes, filosofía, cine, historia, ciencia, economía, etc. También se integra de recopilaciones de algunas secciones de periódicos y revistas de temas como: política, economía, sociedad, cultura, sucesos, entre otros. De igual forma contiene transcripciones de la lengua oral procedente de la radio y televisión **(WB, 15).**

2.4.3 Herramientas de análisis

El análisis de un corpus lingüístico debe contar con un conjunto de herramientas y métodos que son particularmente útiles para la investigación del uso del lenguaje. Entre las herramientas que pueden sernos de utilidad en el análisis de un corpus lingüístico encontramos las concordancias y las listas de palabras. En seguida exponemos las principales características de algunas herramientas de análisis, la forma en que trabajan y cómo pueden ser utilizadas.

2.4.3.1Lista de palabras

Una lista de palabras es una lista de todos los tipos de palabras de un texto, ordenadas por algún criterio. La lista de palabras más común es la lista de frecuencia de palabras, donde las palabras son clasificadas en orden descendente respecto a su frecuencia, de esta forma las palabras más comunes están al principio de la lista y las menos comunes al final.

Una lista de frecuencia de palabras nos puede dar una primera impresión de cómo está conformado un texto. También nos puede ayudar a comparar diferentes textos, mostrando la frecuencia de las palabras en un texto para cotejarlo con la lista de palabras de frecuencia de otro.

Otro tipo de lista de palabras muy usado es el basado en orden alfabético, ya sea desde el inicio de la palabra o desde el final (lista invertida). Este tipo de listas permite analizar las distintas formas de una palabra.

2.4.3.2Concordancias

Como ya mencionamos las listas de palabras pueden ser útiles para darnos una primera impresión del texto, mostrarnos como son utilizadas las palabras e incluso el estilo en que está escrito, pero esto de una manera general. Por ello, resultaría más útil una herramienta que permita ver cada palabra de forma individual, para analizar el contexto y el sentido en el que está siendo utilizada, esta herramienta se llama, generador de concordancias.

El concepto de concordancia ha estado presente casi desde los inicios de los corpus lingüísticos, pero sólo con el avance de las computadoras y las ventajas que estas representan

para el procesamiento y manipulación de datos es que las concordancias han sido realmente aprovechadas. En sus inicios, el tiempo para generar concordancias era tan extenso que sólo se hacían concordancias de textos religiosos o de ciertos trabajos literarios. Sin embargo, ahora es posible crear concordancias de importantes cantidades de textos en pocos segundos utilizando una computadora.

Existen básicamente dos tipos de concordancias, las cuales se refieren a la forma en que son presentadas visualmente[13]:

- **KWIC (keyword in context):** presenta las concordancias en una lista de líneas, donde la palabra clave está en el centro y a los lados el contexto en el que ésta está escrita. Este tipo de concordancias se presenta en una ventana llamada ventana de la concordancia.

  La ventana de la concordancia se refiere a la cantidad de texto que puede acompañar a la palabra sobre la cual se realiza la concordancia y puede ser de tamaño fijo o variable. Una ventana es de tamaño fijo, cuando al presentarse la concordancia el resultado se limita a mostrar cierto número de caracteres o palabras acompañando a la concordancia ya sea, a la izquierda, derecha o ambos lados. Mientras que una ventana de tamaño variable es cuando el resultado de la concordancia se ajusta a una oración o párrafo que acompañan la palabra **(WB, 16)**. La Figura 1 nos ayuda a ilustrar como se presentan una serie de concordancias en una ventana de tamaño fijo.

[13] Si bien la primera versión del CHEM únicamente muestra ventanas de concordancias del tipo KWIC y de tamaño fijo, el equipo de desarrollo planea que la nueva versión del CHEM se permita mostrar ventanas de concordancia de tamaño variable.

Resultados de la búsqueda

#	Siglo		Pal.		Referencia
1	XVI	y no quiere. Soplico a vuestra merçed le haga	**bolver,**	sabiendo vuestra merçed la neçeçidad que ay a	Company C., *DLNE,* México, UNAM, 1994.
2	XVI	çedes, e se tenga consideraçion a que me mandó	**bolver**	a estas partes y me mandó casar y ansj, lo cum	Company C., *DLNE,* México, UNAM, 1994.
3	XVI	{2v} la fama de los treinta que a él yvan, sin	**bolver**	ninguno por ellos, como se acostumbra hazer en	Company C., *DLNE,* México, UNAM, 1994.
4	XVI	tismo rrescibieron, está claro que se avian de	**bolver**	a sus rritos e ydolatrias y, por el consiguien	Company C., *DLNE,* México, UNAM, 1994.
5	XVI	a vuestra majestad, cuya voluntad en mandarlos	**bolver**	a sus tierras se cree que fue presuponiendo se	Company C., *DLNE,* México, UNAM, 1994.
6	XVI	eseo de la nuestra, y algun dia se le antojará	**bolver,**	y hallarse á sus casas y hazienda en pie.\ N	Company C., *DLNE,* México, UNAM, 1994.
7	XVI	geres, porque los hazen luego el rey	**bolver**	a España, o que las traygan. Yo estoy bueno, b	Company C., *DLNE,* México, UNAM, 1994.
8	XVI	rte en esta flota y, dandole Dios salud, ha de	**bolver**	en la misma flota, y creo bendra por esa villa	Company C., *DLNE,* México, UNAM, 1994.
9	XVI	o yo, hallandose por esta causa sin cosa a que	**bolver**	los ojos, y en una tierra donde faltando el vi	Company C., *DLNE,* México, UNAM, 1994.
10	XVI	dose por esta causa sin galardon ni cosa a que	**bolver**	los ojos, y en una tierra donde faltando el vi	Company C., *DLNE,* México, UNAM, 1994.
11	XVI	las que presento, no las a hallado para me las	**bolver.**	\ A vuestra alteza suplico se le pregunte si e	Company C., *DLNE,* México, UNAM, 1994.

Figura 1. Ventana de concordancias de tamaño fijo

- **KWOC (keyword out context):** muestra la palabra clave en el margen de la página, seguida de una oración o párrafo que forma parte del contexto. Por ejemplo, en la Figura 2 se puede observar una concordancia de este tipo para la palabra *pedro* obtenida del CREA.

REAL ACADEMIA ESPAÑOLA

Párrafos (RAE)

Consulta:	*pedro*, en todos los medios, en *CREA*
Resultado:	(Filtrado) 14 casos en 5 documentos.

OBTENCIÓN DE EJEMPLOS

Recuperar | Concordancias | Normal | Clasificación: | Agrupación: | Marcas:

Cómo citar el CORPUS **Párrafos.**

Pantalla: 1 de 1. Ver concordancias

Párrafo nº 1.

Trabajadores. Órgano de la Central de Trabajadores de Cuba, 19/12/2003 : CULTURA PRENSA 5 Trabajadores Digital La Habana 2003 2003 10 402 P

CULTURA

Proclamados los Premios Nacionales de la Música 2003

pedro de la hoz

Seis auténticos pilares de la identidad sonora cubana, Juan Formell, Manuel Duchesne Cuzán, Domingo Aragú, Celina González, Lázaro Ros y Luis Carbonell, se hicieron acreedores de los Premios Nacionales de la Música 2003, máximo reconocimiento que el Ministerio de Cultura otorga a los creadores e intérpretes en esta fértil zona de la cultura insular.

Al proclamar ayer ante la prensa la decisión del jurado, encabezado por Leo Brouwer e integrado por quienes recibieron ese estímulo en las dos convocatorias anteriores, Abel Acosta, viceministro de Cultura y presidente del Instituto Cubano de la Música, expresó que "la selección honra ejemplares trayectorias artísticas que enaltecen a nuestro pueblo".

Entre 77 propuestas evaluadas rigurosamente sobresalieron la extraordinaria vocación renovadora de Formell en la música bailable; el esencial protagonismo de Duchesne Cuzán en la difusión de la vanguardia musical internacional y cubana; la contribución fundacional de Domingo Aragú a la formación y desarrollo de la percusión sinfónica; la dimensión emblemática de Celina González como reina indiscutible de la música campesina; la excepcional voz de Lázaro Ros como portador y promotor de una parte fundamental del patrimonio folclórico; y la subterránea pero imprescindible labor pedagógica y como repertorista de Luis

- -

AÑO: 2003
AUTOR: PRENSA
TÍTULO: Trabajadores. Órgano de la Central de Trabajadores de Cuba, 19/12/2003 : CULTURA
PAÍS: CUBA
TEMA: 04.Música
PUBLICACIÓN: Trabajadores Digital (La Habana), 2003

Figura 2. Ejemplo de concordancia KWOC

2.4.4 Usos o aplicaciones

Un corpus lingüístico resulta importante para conocer qué caracteriza al lenguaje que las personas usamos en diversas situaciones. Con el tiempo, los corpus lingüísticos se han vuelto un recurso imprescindible para la construcción de tecnologías del lenguaje y para el estudio en la ingeniería lingüista. En seguida exponemos algunos usos y aplicaciones de los corpus electrónicos en las áreas como la computación y la lingüística.

2.4.4.1Lingüísticos

Los análisis lingüísticos en un corpus se pueden dividir en dos áreas principales: el análisis de estructura y el análisis de uso del lenguaje. Los análisis de estructura identifican las unidades estructurales del lenguaje y las clases que lo forman (por ejemplo, las clases gramaticales) y describen cómo esas unidades pueden ser combinadas; mientras que los análisis de uso estudian cómo los hablantes y escritores explotan los recursos de su lenguaje.

En la realidad son mucho más comunes los corpus destinados a analizar el uso de la lengua, sin embargo, este tipo de análisis debe ir más allá de un simple conteo de las unidades lingüísticas. La investigación basada en corpus no sólo debe informar de resultados cuantitativos, sino también explorar la importancia de esos hallazgos para el aprendizaje acerca de los patrones de uso del lenguaje.

El uso de un corpus lingüístico nos ayuda, entre otras cosas, a dar seguimiento a factores como la época o la zona geográfica de la lengua. Además, un corpus ayuda en gran medida a evitar las dificultades que pudiera presentar el análisis de grandes volúmenes de datos de diversas variantes de la lengua.

De la diversidad de investigaciones lingüísticas que se pueden realizar sobre un corpus mencionamos las siguientes: detección de neologismos y términos, estudios sobre variación lingüística, análisis sintáctico, 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.

Algunos ejemplos más precisos de los usos lingüísticos de un corpus son:

- En la historia de la lengua, los corpus pueden proporcionar datos referentes a la formación de palabras, a los cambios en su significado, a las diferentes áreas de la utilización de una palabra, así como la introducción de palabras no normativas en la lengua.

- En el campo de la enseñanza de lenguas facilitan la preparación de materiales o ejercicios basados en el uso real de la lengua.

- En el campo de la estilística, un corpus puede ayudar a definir los trazos que caracterizan a los distintos estilos literarios.

- La Real Academia Española utiliza sistemáticamente el CORDE para documentar palabras, calificarlas de anticuadas o en desuso, saber el origen de algunos términos, su tradición en la lengua, las primeras apariciones de las palabras, entre otras funciones.

- En la psicolingüística, se utilizan para estudiar patologías del lenguaje y del habla.

- En la sociolingüística se pueden utilizar corpus para analizar datos como la clase social, el sexo o el nivel cultural de un hablante.

- Los corpus hablados resultan útiles para el estudio experimental del habla. Por ejemplo, un corpus que contiene recopilaciones orales de estudiantes de lengua extranjera ayudará a conocer la interferencia entre la primera y la segunda lengua en todos los niveles del análisis lingüístico, ayudando a desarrollar estrategias de comunicación para los alumnos.

2.4.4.2Computacionales

Los corpus electrónicos son una herramienta primordial para muchos tipos de investigaciones, principalmente lingüísticas, porque proporcionan bases mucho más reales para el estudio de la lengua que algunos métodos intuitivos tradicionales. Una de las áreas en las que más ha impactado el uso de estos corpus es la computación.

Algunos ejemplos del uso de un corpus electrónico son los siguientes:

- Un corpus electrónico puede ayudar para la creación de herramientas lingüísticas automatizadas. Por ejemplo: los diccionarios-máquina que se usan para la corrección de textos automatizados así como para la traducción automática.

- En cuanto a las bases de datos orales, generadas a partir de un corpus, proporcionan datos importantes para el modelado de herramientas que sirvan para la conversión de texto en habla (síntesis de voz) y son esenciales para el entrenamiento y la validación de los sistemas de reconocimiento y de diálogo en entornos de comunicación persona-

máquina. Por ejemplo una oferta de servicios telefónicos automatizados o ayuda a personas con discapacidades.

3. El Corpus Histórico del Español en México (CHEM)

3.1 Antecedentes

Ya en la introducción de esta tesis, nos referíamos a los antecedentes del CHEM, por lo que en este capítulo nos enfocaremos a resumir la información dada. La Universidad Nacional Autónoma de México, como decíamos, ha permitido el surgimiento de numerosos institutos con la intención de generar conocimientos y tecnologías útiles para el desarrollo de nuestro país. Estos institutos a su vez pueden integrarse por grupos de investigación en áreas más específicas.

Uno de los institutos de investigación de la UNAM es el Instituto de Ingeniería, el cual contiene entre sus grupos de investigación al Grupo de Ingeniería Lingüística. El GIL ha venido desarrollando investigación para la construcción de corpus lingüísticos electrónicos desde hace ya un largo periodo y destaca que ha puesto especial interés y apoyo a la realización del CLI y del CSMX, así como al propio CHEM.

El Corpus histórico del español en México forma parte de los proyectos patrocinados por el PAPIIT de la Dirección General de Asuntos del Personal Académico de la UNAM y tiene el número de proyecto IN402008. El objetivo del PAPIIT es impulsar el desarrollo de proyectos de investigación básica, aplicada y multidisciplinaria de alta calidad en diversas áreas del conocimiento.

Un corpus puede ser utilizado en diversos tipos de investigaciones, por ejemplo, aquellas relacionadas con la inteligencia artificial, el procesamiento del lenguaje natural, extracción y recuperación de información y minería de textos, o incluso investigaciones que involucren alguno de los niveles de estudio del lenguaje. El propósito del CHEM es ayudar a estas investigaciones, proporcionando las herramientas y materiales para su logro.

Cabe mencionar que el CHEM es desarrollado a partir de materiales que lingüistas, filólogos e historiadores rescatan y consideran clave para representar el español de México entre los siglos XVI y XIX. En el CHEM participan además estudiantes y académicos que se encargan de construir herramientas y tecnologías para el análisis de esos materiales.

3.2 CHEM

La existencia del CHEM constituye un repositorio documental, patrimonio cultural, útil para realizar investigaciones de diversos tipos, tanto en lingüística, como en desarrollos para tecnologías del lenguaje. Además, se prevé que en poco tiempo contribuya en áreas como las humanidades o las ciencias sociales, y especialmente en la Ingeniería lingüística. También, el CHEM podrá beneficiar a aquellos investigadores interesados en el dialecto español mexicano.

El objetivo principal del CHEM consiste en "Compilar una colección de documentos históricos escritos en la Nueva España y el México independiente y las herramientas para analizarlos; en otras palabras, constituir un *corpus* diacrónico para el español de México (siglos XVI al XIX), que pueda ser consultado mediante herramientas computacionales en una interfaz de Internet" **(WB, 17)**. Lo anterior, permitiría que se puedan hacer búsquedas de estructuras gramaticales específicas ahorrando tiempo y esfuerzo en investigaciones diacrónicas.

Al momento de la realización de esta investigación, está disponible en línea una primera versión el CHEM que permite hacer uso de un generador de concordancias para hacer análisis dentro del corpus. La versión se encuentra en la página: http://www.iling.unam.mx/chem/.

Entre los objetivos particulares del CHEM podemos resaltar dos **(WB, 18)**:

- El primer objetivo es diseñar y desarrollar herramientas para hacer exploraciones específicamente diacrónicas del corpus.
- El segundo tiene que ver con diseñar el mapa del corpus por cada siglo (XVI, XVII, XVIII y XIX), tomando en cuenta géneros literarios (prosa, poesía, ensayo, etc.) y temáticos (literatura, gobierno, religión, ciencia, etc.), tipos de textos (libros, artículos, periódicos), autores prominentes, colecciones y archivos disponibles.

3.3 Arquitectura del CHEM

Parte fundamental del CHEM ha sido el diseño de su arquitectura, la cual está compuesta por herramientas que permiten analizar los materiales contenidos en el corpus (como el generador de n-gramas y el generador de concordancias) así como del proceso de

incorporación de los documentos al CHEM[14]. En los siguientes apartados haremos una breve descripción de cada una de las partes que conforma la arquitectura del CHEM, tomando como base para ello el documento "Arquitectura del Corpus Histórico del Español de México (CHEM)" **(Medina y Méndez, 2006: 1-6)**.

Aunque destacamos que la administración de usuarios y documentos del corpus forman parte de la arquitectura del CHEM, los describiremos en capítulos posteriores debido a que son parte primordial del desarrollo de esta tesis.

3.3.1 Generador de n-gramas

Para comenzar con la descripción de la arquitectura, indicamos que el CHEM contará con una base de datos relacional, que entre otras cosas permite la construcción de tablas de n-gramas, principalmente de unigramas (una palabra), mediante el generador de n-gramas. El generador de n-gramas se encarga de tokenizar[15] los textos y conformarlos en n-gramas. Para cada grama, un indexador de archivos obtiene su posición en cada documento, es decir, los bytes necesarios para acceder a éste en el texto; al mismo tiempo, obtiene su descripción lingüística asociada en tres niveles: nivel fonológico (transcripción fonológica basada en un alfabeto fonológico), de lema (palabra de diccionario asociada al grama) y POS (la categoría gramatical asociada al grama).

Es importante mencionar que también se obtienen las frecuencias de aparición de cada grama en los diferentes documentos.

[14] La arquitectura del CHEM ha tomado como base el método utilizado por Mark Davis para la infraestructura de su Corpus del Español (disponible en: www.corpusdelespanol.org*),* pero con mejoras como el etiquetado en el corpus mismo. Mark Davies, por su parte, realiza el etiquetado dentro de la base de datos relacional del corpus.

[15] Como tal, la palabra "tokenizar" no existe en el diccionario de la Real Academia Española; sin embargo, es el término en inglés para el proceso de separación de palabras y signos de puntuación de un corpus.

3.3.2 Generador de concordancias

El generador de concordancias es una herramienta que hace posible el análisis del corpus lingüístico. Está compuesto por una serie de clases en lenguaje java, que toma la petición del usuario, que puede ser una o varias palabras, y regresa una ventana (contexto) con un determinado número de palabras a la izquierda y a la derecha de la palabra que originalmente se buscó, lo anterior tantas veces como se encuentre la palabra en los documentos[16].

El generador de concordancias básicamente actúa de la siguiente forma:

1. Toma la petición el usuario.
2. Ejecuta el SELECT correspondiente sobre las tablas de n-gramas.
3. Obtiene los documentos y las posiciones en los archivos de cada palabra buscada.
4. Con el resultado de la consulta anterior, accede a los textos y se recupera la ventana de caracteres.
5. La ventana de caracteres pasa a la interfaz de salida que da formato a la concordancia.

Parte de esta tesis es utilizar el generador de concordancias asociado a un documento completo. Para ser más específicos, cada palabra del documento tendrá un enlace HTML que permitirá la recuperación de sus concordancias, por lo que cada una será susceptible de recuperarse en su contexto textual.

3.3.3 Proceso de incorporación de documentos

En lo que se refiere a los documentos, estos son el componente principal del CHEM. Son materiales, como hemos mencionado antes, pertenecientes a los siglos XVI al XIX que han sido proporcionados por lingüísticas y filólogos de diversas instituciones.

Cada documento contiene información bibliográfica, es decir su fuente documental, con información sociolingüística como región originaria del hablante, género, etc.; y

[16] Precisamos que en la primera versión del CHEM, el generador de concordancias regresaba los resultados de las búsquedas con caracteres a la derecha e izquierda de los mismos y no con palabras.

discursiva, que se refiere al tipo de documento. Los documentos están estructurados y anotados en lenguaje XML por lo que la información anterior está contenida en un encabezado XML.

Mantener los documentos del CHEM como una base de datos XML permite mantener la flexibilidad del corpus en lo que respecta a la incorporación de nuevos documentos así como a su fácil transformación a otros formatos. Los documentos también utilizan un esquema XML (schema XSD), que permite corroborar que la información está completa y bien estructurada.

A continuación presentamos un ejemplo del proceso para la incorporación de nuevos documentos al CHEM.

- En primer lugar, un filólogo o lingüista identifica el documento original, el cual se puede encontrar impreso como el de la Figura 3.

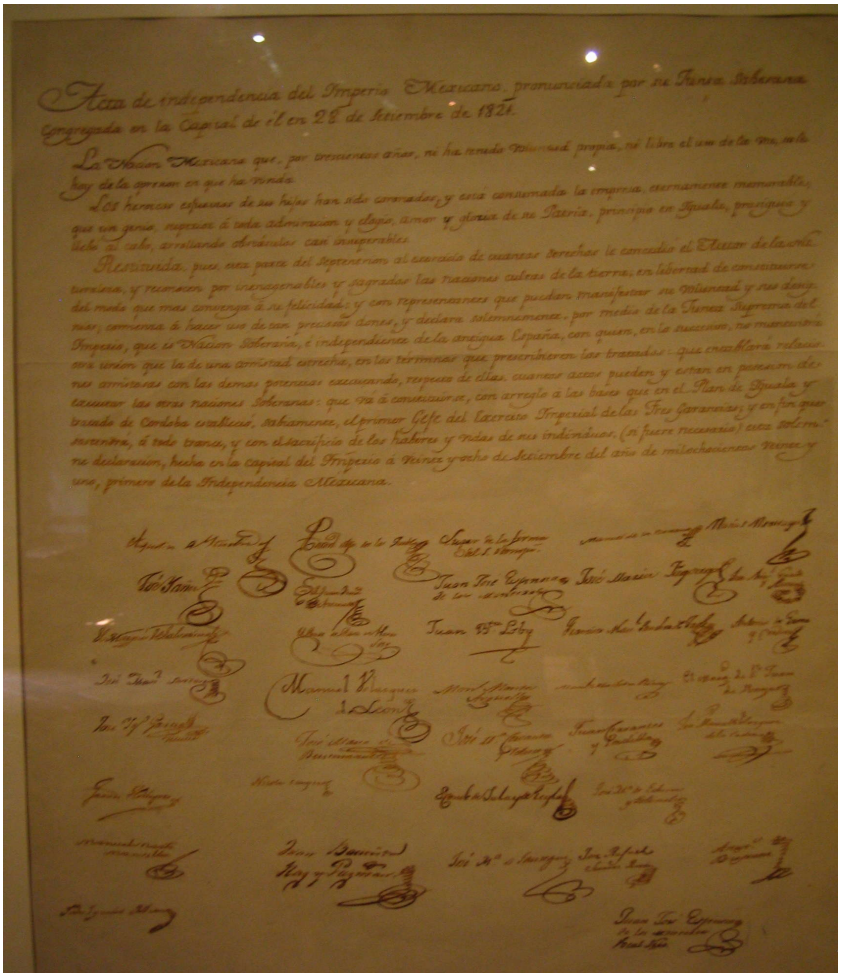


Figura 3. Documento original impreso

- Después, el documento es transcrito a formato electrónico, es decir, se convierte en un documento electrónico como el siguiente.

Acta de independencia del Imperio Mexicano,

pronunciada por su Junta Soberana

congregada en la Capital de él en 28 de Setiembre de 1821./

La Nacion Mexicana que, por trescientos años, ni ha tenido voluntad propia, ni libre el uso de la voz, sale/ hoy de la opresion en que ha vivido./

Los heroicos esfuerzos de sus hijos han sido coronados y esta consumada la empresa eternamente memorable/ que un genio, superior á toda admiración y elogio, amor y gloria de su Patria, principio en Iguala, prosiguio y/ llebo al cabo, arrollando obstáculos casi insuperables./

Restituida, pues, cada parte del Septentrion al exercicio de cuantos derechos le concedió el Autor de la na-/ turaleza, y reconocen por inenagenables y sagrados las naciones cultas de la tierra, en libertad de constituirse/ del modo que mas convenga a su felicidad, y con representantes que pueden manifestar su Voluntad y sus desig-/ nios, comienza á hacer uso de tan preciosos dones, y declara solemnemente, por medio de la Junta Suprema del/ Imperio, que es una Nación Soberana é independiente de la antigua España, con quien, en lo sucesivo, no mantendrá/ otra unión que la de una amistad estrecha, en los términos que prescribieren los tratados; que entablará relacio-/ nes amistosas con las demas potencias, executando, respecto a ellas, cuantos actos pueden y estan en posesion de/ executar las otras naciones Soberanas; que va á constituirse, con arreglo á las bases que en el Plan de Iguala y/ Tratado de Córdoba estableció, sabiamente, el Gefe del Exercito Imperial de las Tres Garantías, y en fin que/ sostendrá, á todo trance, y con el sacrificio de los haberes y vidas de sus individuos, (si fuere necesario) esta solem-/

Figura 4. Vista preliminar de un documento electrónico

- Por último, el documento electrónico es etiquetado con el lenguaje de marcas XML. Una vez que el documento está en formato XML, es susceptible de ser agregado al CHEM para brindar la posibilidad de ser analizado, por ejemplo, por el generador de concordancias. El documento XML se ve como la Figura 5.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet type="text/xsl" href="estilo.xsl"?>
<documento xmlns="http://www.ii.unam.mx"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.ii.unam.mx esquema.xsd">
<encabezado id="1821092801" permisos="todos">
<titulo>Acta de independencia del Imperio Mexicano</titulo>
<parámetros generoLiterario="prosa" registro="estándar">
<corpus>
<CHEM
</corpus>
</parametros>
<hablante id="2"/>
<referencia id="chem1">
<lugar>ciudad de México</lugar>
<fechaOriginal>28 de septiembre de 1821</fechaOriginal>
</referencia>
<imagen origen="internet" ancho="400" alto="532">Independencia-mx-acta.jpg</imagen>
<responsables>
<transcriptor nombres="Alfonso" apellidos="Medina Urrea" fecha="noviembre 2007"/>
<etiquetador nombres="Alfonso" apellidos="Medina Urrea" fecha="noviembre 2007"/>
<revisor nombres="" apellidos="" fecha="noviembre 2007"/>
</responsables>
</encabezado>
<cuerpo tipoFuente="Espinosa">
<bastardillas>
<seccion>
<tituloSecc>
<g>Acta</g><e/>
<g>de</g><e/>
<g>independencia</g><e/>
<g>del</g><e/>
<g>Imperio</g><e/>
<g>Mexicano</g><r c="Fc">,</r><e/>
</tituloSecc>
</seccion>
</bastardillas>
<objeto tipo="rúbricas">
<nota llamada="a" notaEditor="Los miembros de la Regencia del Imperio: Agustín de Iturbide, Presidente.
Juan O'Donojú, Segundo regente. Manuel de la Bárcena, Tercer regente. José Isidro Yañez, Cuarto regente.
Manuel Velásquez de León, Quinto regente."/>
</objeto>
</cuerpo>
</documento>
```

Figura 5. Vista preliminar de documento con formato XML.

3.3.4 Base bibliográfica XML

Actualmente XML se ha posicionado como uno de los estándares de intercambio de información y estructuración de documentos más utilizado en el mundo. Es por ello y entre otras cosas, que el total de textos del corpus está contenido en archivos XML, con esto se hace posible que los documentos puedan ser explotados mediante lenguajes de consulta y que además, como mencionamos en el apartado anterior, se permita mantener la flexibilidad del corpus en lo que respecta a la incorporación de nuevos textos.

4. Administración del CHEM

En este capítulo especificaremos en qué consistirá la administración en la nueva versión del CHEM, tanto de usuarios como de documentos, pero primero hacemos referencia a los problemas del registro de usuario de la primera versión para luego describir los diferentes tipos de usuarios que se contemplarán en el CHEM. Posteriormente detallamos qué diferencia a cada uno de los usuarios, sus permisos y privilegios. Por último, tratamos el tema de los documentos, señalando los diferentes permisos a los que pueden estar asociado cada uno y describiendo en qué consisten estos.

4.1 Administración de Usuarios

Ya mencionábamos, al principio de este documento, que el CHEM carecía de una administración de usuarios. Lo anterior, porque el tamaño del corpus no lo requería debido a que únicamente se había desarrollado una herramienta para la explotación del corpus y además los documentos completos no estaban disponibles para ser visualizados en línea. Los siguientes aparatados son el resultado del análisis a la necesidad de de una administración de usuarios.

4.1.1 Registro de Usuario

En el planteamiento del problema señalábamos los inconvenientes del registro de usuario, los cuales son: en primer lugar, que se consiente que los usuarios omitan ciertos datos y, en segundo lugar, que esos datos pueden ser incorrectos. A continuación describiremos el funcionamiento del registro de usuario de la primera versión del CHEM con la finalidad de indicar cómo es que los errores mencionados se presentan.

El registro de usuario sigue los siguientes pasos:

- El primer paso consiste en aceptar el Acuerdo general para usuarios del Corpus Histórico del Español de México.
- Después se presenta al usuario un formulario que solicita datos personales como: nombre, apellidos, correo electrónico, contraseña y fecha de nacimiento; datos de procedencia como: país e institución; y la ocupación así como también se le solicita la

elección de una pregunta secreta y la respuesta de la misma para permitirle posteriormente la recuperación de su contraseña.

- Por último, una vez que el usuario ha aceptado los datos, éstos le son mostrados para confirmar su registro.

Después de realizar los pasos anteriores el usuario puede con su correo electrónico y contraseña acceder al generador de concordancias.

Los pasos anteriores encierran algunos inconvenientes, por ejemplo que en ningún momento se hace alguna validación en cualquiera de los datos que el usuario proporciona, así que resulta muy fácil realizar un registro que incluya únicamente el nombre, correo electrónico o contraseña y omita los demás datos. Además, a falta de un método que revise la estructura de los datos, un usuario puede proporcionar datos que no tienen nada que ver con lo que se le está solicitando.

Otro problema, derivado de los anteriores, es que un usuario puede proporcionar datos que resulten ser instrucciones de consulta a la base de datos, lo cual tendría repercusiones importantes a la integridad de la información que pueda contener el CHEM.

4.1.2 Análisis e identificación de Usuarios

Para la nueva versión del CHEM se han identificado a dos principales grupos de usuarios:

1. Usuarios registrados.
2. Usuarios anónimos.

Un usuario registrado será aquel que haya proporcionado datos como: nombre, correo electrónico, contraseña, fecha de nacimiento, país, institución a la que pertenece y ocupación a un formulario de registro del propio CHEM; además de que ha aceptado el Acuerdo General para Usuarios del CHEM y ha dado de alta y confirmado que su cuenta de correo electrónico es auténtica a través de un enlace (link) que se le envía a su correo electrónico[17].

[17] En el apartado 4.1.4 se explica el tema de confirmación de registro por medio de un correo electrónico.

Dentro del grupo de usuarios registrados se podrán distinguir a tres subtipos de usuarios que a continuación listamos:

- Administrador
- Usuario avanzado
- Usuario básico

Destacamos al administrador, porque es quien se encargará de asignar privilegios a los usuarios, así como de mantener el control de los mismos, acceder al libro de visitas y consultar las estadísticas que se generen como resultado del uso del CHEM.

Cabe señalar también, que cuando un usuario se registre, en principio será reconocido únicamente como un usuario registrado de tipo básico, posteriormente si el administrador del CHEM lo cree conveniente, podrá actualizar el registro del usuario para darlo de alta como un usuario registrado de tipo avanzado. En los siguientes apartados ampliaremos la descripción de los usuarios para resaltar las diferencias que existen entre cada uno de ellos.

Por su parte, un usuario anónimo será aquel que sin iniciar previamente una sesión haga uso de los materiales del CHEM. En este punto aclaramos que al ser el CHEM un producto resultado de investigaciones realizadas dentro de la UNAM y con apoyo de la misma, se ha convenido tratar a aquellos usuarios anónimos que accedan a la aplicación del CHEM dentro de alguna de las instalaciones de Ciudad Universidad como usuarios registrados de tipo básico[18]. Lo anterior concede algunos privilegios a la comunidad universitaria sin tener que pasar por el proceso de registro.

Se hace evidente entonces que un usuario anónimo también puede ser subdivido en otros dos tipos de usuarios que a continuación listamos:

- Usuario anónimo externo a la UNAM.
- Usuario Anónimo dentro de la UNAM.

[18] Para lograr esta distinción se planea estudiar algunos métodos que permitan identificar la dirección IP de los usuarios. En el siguiente apartado se expondrán algunas opciones para lograr este objetivo.

Es importante recalcar, de acuerdo a las especificaciones anteriores, que un usuario no deberá obligatoriamente estar registrado para poder hacer uso del CHEM; sin embargo, el realizar su registro se le ofrecerá la oportunidad de explotar de una mejor forma los contenidos del corpus. Un usuario registrado adquirirá privilegios que le agregan valor a los resultados de sus investigaciones teniendo un mayor acceso a los contenidos del corpus.

Finalmente, notamos que podría parecer elevado el número de tipos de usuarios del CHEM después del desarrollo de esta tesis, dado que únicamente se contará con dos herramientas: el generador de concordancias y el visor de documentos completos (tema que trataremos más adelante); pero esto tiene su justificación. El proyecto del CHEM, al ser un proyecto de investigación, contempla la construcción e integración a corto o largo plazo de más herramientas de explotación de corpus y todas ellas deberán ser administradas.

4.1.2.1Estudio de métodos para la obtención de la dirección IP

Para poder identificar a aquellos usuarios que hacen uso del CHEM dentro de la UNAM, se acordó utilizar un método que permitiera identificar la dirección IP de los usuarios, con el fin de diferenciarlos y con esto concederles algunos permisos sobre los materiales del corpus.

Dado lo anterior, nos dimos a la labor de buscar un método que nos permitiera cumplir con el objetivo, a continuación exponemos el resultado del estudio a algunos métodos para la obtención de la dirección IP.

- Geobytes: este método consiste en que una página de internet captura la dirección IP del usuario que ingresa a la misma. Para lograr aplicar dicho método al CHEM, se hace uso de un script, el cual regresa una serie de variables que contienen la información de la IP y el nombre del país desde el cual la página está siendo visitada y consultada[19].

[19] Aclaramos, que como la página de Geobytes existen muchas otras dentro de la WEB que realizan la misma tarea (obtener la dirección IP), por lo que únicamente nos dedicamos a exponer la que posiblemente será integrada a la aplicación de consulta. Otra página que provee el mismo servicio es: Maxmind.

- Applets: un applet es un script con código java que se ejecuta del lado del servidor, en este caso, nos encontramos con un applet que puede capturar la IP del usuario, sin embargo esta IP resulta de poca utilidad debido a que la información que brinda no es tan precisa como se requiere para el CHEM.

4.1.3 Permisos de Usuarios

Como mencionamos en la introducción de este documento, uno de los problemas a los que se busca dar solución es la administración de usuarios del CHEM. Por lo que, con la finalidad de tener un mejor control de quienes usan los contenidos del corpus, es que se ha restringido el uso de los materiales y herramientas. Dependiendo del grado de confianza que se pueda tener sobre cada uno de los diferentes usuarios se acordó asignar diferentes permisos. Lo anterior con el objetivo de satisfacer las necesidades de seguridad que requieren los documentos que forman parte del CHEM.

En este apartado enumeramos los diferentes tipos de usuarios, agregando también el detalle de los permisos correspondientes a cada uno de los documentos que puede visualizar, el tamaño de ventana de concordancias, y los documentos de los cuales pueden extraer concordancias.

En una sección posterior describiremos en qué consiste cada uno de los permisos de los documentos, por lo pronto sólo explicamos los permisos del usuario. Antes aclaramos que el tamaño de la ventana de concordancias se refiere a la posibilidad de aumentar o reducir el número de palabras que acompañen a la concordancia, tanto a la izquierda como a la derecha. También, los documentos que puede visualizar son aquellos que el usuario puede ver como texto completo, sin necesidad de usar el generador de concordancias. Por último, las concordancias dependen del permiso que se tiene sobre los documentos para buscarlas, es decir, el resultado de las concordancias sólo será producido a partir de los documentos permitidos.

Para poder facilitar la comprensión de la asignación de permisos, ponemos a continuación dos tablas en las que se listan éstos. Hemos decido presentarlas con base en la distinción entre los dos principales grupos de usuarios: registrados y anónimos.

Tabla 1. Asignación de permisos a usuarios registrados

Usuario Registrado	Tamaño de ventana de concordancias	Documentos que puede visualizar	Concordancias
Administrador	Variable	Todos los documentos: Sin permiso Con permiso En trámite Parciales Sin permiso/libre	Todos los documentos: Sin permiso Con permiso En trámite Parciales Sin permiso/libre
Avanzado	Variable	Con permiso En trámite Sin permiso/libre	Con permiso En trámite Parciales Sin permiso/libre
Básico	Variable	Con permiso	Con permiso

Tabla 2. Asignación de permisos a usuarios anónimos

Usuario Anónimo	Tamaño de ventana de concordancias	Documentos que puede visualizar	Concordancias
Usuario dentro de la UNAM	Variable	Con permiso	Con permiso
Usuario externo a la UNAM	Fijo	Con permiso	Con permiso

4.1.4 JavaMail para confirmación de registro y recuperación de contraseña

Anteriormente señalábamos que, una vez que el usuario proporciona sus datos, es necesario que active y confirme su cuenta por medio de un enlace que le es enviado vía correo electrónico a la dirección que utilizó para registrarse. En este apartado precisaremos el proceso de envío de dicho correo electrónico y la forma en que este actúa para dar de alta la nueva cuenta de usuario.

Uno de los problemas con los que nos encontramos, asociado al envío de correo electrónico, es que normalmente para ello se necesita de un servidor especial capaz de proporcionar dicho servicio. Lo anterior implica por ende dedicar más tiempo a la configuración del servidor.

Afortunadamente, luego de revisar algunos métodos para el envío de correo electrónico, nos encontramos con una Interfaz de Programación de Aplicaciones (API, por sus siglas en inglés)[20]. Esta API es desarrollada con el lenguaje de programación Java y distribuida por Sun Microsystems[21]. La API se llama JavaMail y provee al marco de trabajo (framework) de una plataforma y protocolos independientes que permiten la construcción de correo y aplicaciones para envío y recepción de mensajes **(WB, 19)**.

JavaMail también trae consigo el cumplimiento de un requerimiento solicitado por el GIL, el de tratar de encontrar soluciones tecnológicas que no impliquen cambios en el servidor del grupo, por lo que resulta más que adecuada.

Como lo mencionábamos, un correo es enviado cada vez que un usuario proporciona sus datos cuando desea registrarse en el CHEM, pero ¿por qué decidimos enviar un correo electrónico para concluir el registro? La respuesta es la siguiente: enviando un correo electrónico como forma de concluir el registro se podría, en primer lugar, verificar que la

[20] Una API es un conjunto de bibliotecas que contienen métodos y procedimientos para ser usados por diferentes aplicaciones brindando una forma de comunicación.

[21] JavaMail se encuentra disponible para su descarga en la siguiente página: http://java.sun.com/products/javamail/.

cuenta de correo existe y que además es una persona quien está haciendo el registro[22]. Lo anterior además, también brinda la posibilidad de optimizar la base de datos del CHEM al activar únicamente las cuentas de aquellos usuarios que concluyan el registro. La estructura de correo electrónico que permitirá la activación de la cuenta del usuario se mostraría de forma similar a la Figura 6.

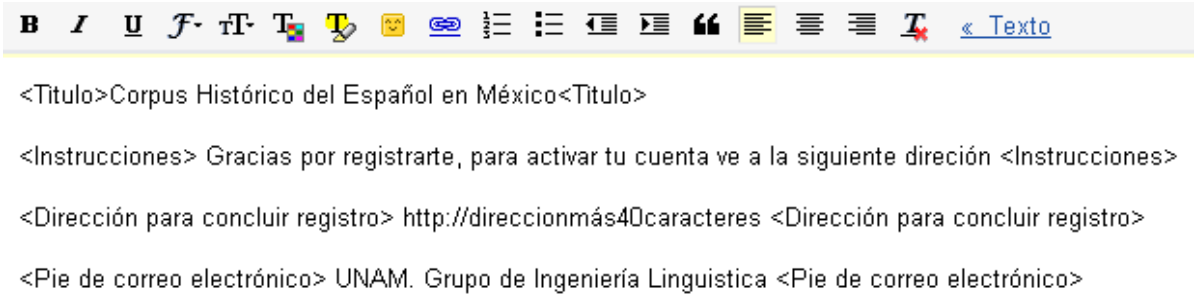


Figura 6. Boceto de correo electrónico para confirmación y activación de cuenta

De igual forma se prevé que algún usuario tenga la necesidad de recuperar su contraseña en caso, por ejemplo, de que la haya olvidado, para esto, también utilizaremos JavaMail ya que cabrá la posibilidad de enviar un correo a aquellos usuarios que requieran recuperar su contraseña. El boceto de la estructura de este correo electrónico se ilustra a continuación con la Figura 7.

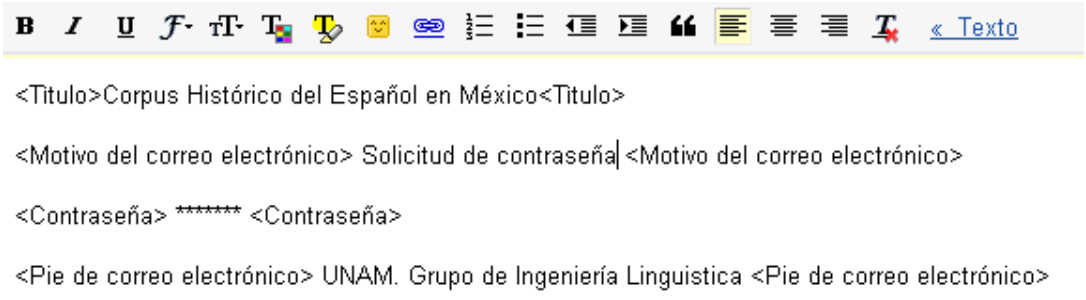


Figura 7. Boceto de correo electrónico para recuperación de contraseña

[22] Hoy en día existen programas que por distintas razones, a veces de forma malintencionada, hacen registros automáticos con datos ficticios.

En el capítulo asignado para la integración de esta investigación definimos el aspecto final de ambos correos electrónicos así como la forma en que esta API, JavaMail se construye para poder funcionar.

4.2 Administración de Herramientas

Como hemos señalado, para la nueva versión del CHEM se prevé la creación e integración de nuevas herramientas por lo que será necesario brindar algunas posibilidades para que por ejemplo un usuario administrador, pueda gestionarlas de manera sencilla.

Las herramientas que se contempla que contenga la nueva versión son: el generador de concordancias, el generador de estadísticas (éste trabajo unido al generador de concordancias), el visor de documentos y el libro de visitas. Por lo que será necesario crear una base de datos capaz de administrar los permisos de las nuevas herramientas, así como su disposición para con los usuarios. También se requerirá de una base de datos que permita la creación de nuevas herramientas para el análisis de corpus que afecte directamente a los documentos que estén relacionados con ellas.

En la parte de integración de esta tesis, exponemos la forma en que podrán ser creadas nuevas herramientas, así como la manera en que los usuarios podrán tener acceso a ellas a través de los permisos.

4.3 Administración de Documentos

Los documentos son el componente principal del CHEM, pero es importante mencionar que aunque en la primera versión del CHEM se contaban con 320 textos provenientes de los *Documentos Lingüísticos de la Nueva España,* editados por la Dra. Concepción Company del Instituto de Investigaciones Filológicas de la UNAM, ninguno de ellos estaba asociado a un nivel de permiso de documento, es decir, no estaba establecido si se podían leer o procesar por cualquier usuario. Sin embargo, para la primera versión esto no era necesario, debido a que en esta versión del CHEM no se incluye la posibilidad de consultar y analizar los materiales en su totalidad (texto completo), únicamente se puede ver el resultado del generador de concordancias, derivado de la búsqueda del usuario.

Por lo anterior, uno de los objetivos que busca cumplir esta tesis es el que tiene que ver con permitir que determinados usuarios tengan acceso parcial o total a los documentos

del corpus, con la intención de que no sólo puedan generar concordancias, sino también para que puedan ver los documentos completos para brindar la oportunidad de realizar análisis lingüísticos y filológicos.

Para logar el objetivo es necesario agregar al CHEM funciones que permitan llevar a cabo la administración de los documentos. Lo anterior implica, debido a la importancia y permisos que el Instituto de Ingeniería tiene sobre los documentos, asociar cada material a un nivel de permiso, esto es, que los documentos pueden tener diferentes permisos y sobre esto se identifica qué documentos pueden ser vistos por qué tipo de usuarios. El nivel de permiso, por consiguiente, brindará mayor seguridad a los materiales, y a los usuarios la oportunidad de ver el texto de los documentos por completo[23].

A continuación, presentamos el resultado del análisis de los diferentes tipos de permisos asociados a los documentos.

4.3.1 Permisos de Documentos

Uno de los criterios más importantes para poder procesar y publicar los documentos contenidos en el CHEM, ha sido el que tiene que ver los derechos de autor, atendiendo a los permisos que se tiene sobre cada documento y al objetivo de proteger tanto los derechos de propiedad intelectual del material como a los propios materiales. Para poder brindar la posibilidad de manejar la seguridad y disponibilidad de los documentos, convenimos categorizar a los mismos de la siguiente manera:

- Documentos con permiso
- Documentos sin permiso
- Documentos con permiso en trámite
- Documentos que no requieren permiso
- Documentos con permisos parciales

[23] Evidentemente los diferentes usuarios tienen diferentes permisos sobre los documentos. Más adelante precisamos la relación entre usuarios y documentos.

Para explicar estos tipos de permisos comenzamos señalando que un documento tiene permiso cuando ha habido un acuerdo previo, explícito y por escrito del titular de los derechos. Las diferentes formas en las que el proyecto se hace de los permisos para usar los materiales consiste en un acuerdo para la sesión de derechos de uso, el cual se puede presentar de las siguientes formas: por medio de cartas de autorización o por convenios que se crean con otras instituciones para explotar sus materiales.

El proyecto hasta ahora ha logrado realizar convenios con instituciones como la Casa de las Humanidades y el Instituto de Investigaciones Filológicas; y ha obtenido cartas de permisos de profesores y alumnos de la Facultad de Filosofía y Letras, y la Universidad Autónoma Metropolitana, entre otros. Lo anterior debido a que los materiales del corpus no pueden perder, por el hecho de ser difundidos a través de Internet, su protección legal.

En cuanto a los documentos que tienen su permiso en trámite, por cuestiones de respeto a los derechos de autor y seguridad de los mismos, sólo son utilizados para generar concordancias y no para ser publicados en el CHEM y ser leídos completamente.

Los documentos que no requieren permiso son un caso especial, estos documentos pueden ser transcripciones de algún material en edición facsimilar, por lo que se vuelven ediciones propias del proyecto y deja de ser necesario que tengan algún permiso. También podemos decir que un documento no requiere permiso cuando se trata de obras públicas de índole general como es el caso de la Constitución de los Estados Unidos Mexicanos.

Por su parte, los documentos asociados a permisos parciales son todos aquellos documentos que pueden ser procesados únicamente para la generación de concordancias.

En cualquiera de los casos anteriores, el CHEM contempla agregar las referencias a las fuentes de cada uno de los textos, señalando los datos bibliográficos pertinentes. Por último, mencionamos que una vez que el corpus esté construido también se protegerá como propiedad intelectual y se reconocerá el trabajo del equipo (diseñadores, programadores, transcriptores, etiquetadores, digitalizadores, etc.) así como el apoyo de los patrocinadores.

4.3.2 Visor de documentos

Una vez resuelto el tema de los permisos de los usuarios y los documentos, es momento de orientar la investigación hacia la búsqueda de una herramienta que permita mostrar los documentos completos en la versión en línea del CHEM. Una herramienta que deba además brindar la posibilidad de construir documentos, relativamente ligeros, que puedan ser cargados rápidamente en Internet, manteniendo las referencias y estilos de cada documento XML que conforma el corpus.

En primer lugar, indicamos que aunque existen diversas formas para mostrar documentos en Internet, el equipo acordó mostrar los documentos como PDF (Portable Document Format), debido a su facilidad de construcción, a la compatibilidad con los clientes y a las medidas de seguridad que estos pueden contener, por lo que las herramientas que se analizarán están enfocadas hacia la construcción de documentos con formato PDF[24].

A continuación exponemos las diferentes herramientas que se cotejaron para ayudarnos a visualizar los documentos en internet dentro de la aplicación del CHEM.

iText

Es una API, es decir, un conjunto de librerías, que permiten generar un PDF dinámicamente utilizando código Java. iText por lo tanto no es un programa como Acrobat, orientado al usuario final, en lugar de eso, iText se integra en las aplicaciones para automatizar el proceso de creación y manipulación de documentos PDF **(WB, 20)**.

iText resulta muy útil para desarrolladores que quieren generar documentos de sólo lectura que pueden contener texto, imágenes, tablas o listas. También para quienes desean

[24] También se contempló mostrar los documentos utilizando paginación o mostrándolos dentro de un documento HTML, sin embargo, cualquiera de las dos opciones resulta peligrosa ya que la única forma de proteger su contenido es mediante Java Script. Aunque Java Script permite proteger los contenidos de ser copiados, es sabido que cualquier persona puede deshabilitar Java Script en su navegador de una forma muy sencilla. Si esto sucediera, el contenido de los documentos podría ser copiado o impreso, lo cual es precisamente lo que se quiere evitar.

realizar manipulaciones específicas sobre documentos PDF existentes. iText se puede usar para las siguientes tareas:

- Generar documentos dinámicos a partir de archivos XML o bases de datos.
- Añadir marcadores, números de página, marcas de agua, etc.
- Concatenar o manipular páginas con PDF.
- Agregar firma digital a un archivo PDF.

Entre los casos en que esta API resulta útil se encuentran los siguientes:

- Cuando debido al tiempo o al tamaño el PDF no puede ser construido manualmente.
- Cuando el contenido del documento se basa en la entrada del usuario.
- Cuando el contenido del PDF necesita estar en un servidor en el ambiente WEB.

Sin embargo, aunque pareciera que iText ofrece la posibilidad de realizar diversas tareas, su documentación es pobre y confusa.

Flash Paper 2

Permite convertir cualquier documento, que pueda imprimirse, en un archivo SWF o PDF. Ambos tipos de archivos, SWF y PDF, pueden además ser embebidos en una página WEB como cualquier objeto.

Flash Paper 2 permite la transformación de archivos en documentos PDF seguros y compactos que pueden ser compartidos en cualquier sitio WEB **(WB, 21)**. Para los documentos SWF también brinda la posibilidad de que los documentos puedan ser seleccionados, copiados, pegados e impresos a través de su interfaz gráfica; así como el reconocimiento e inclusión de enlaces. Esta herramienta también incluye un cuadro de texto para realizar búsquedas dentro del documento **(WB, 22)**.

Los documentos que son creados por esta herramienta se abren dentro de una página Web, eliminando así la necesidad de tener una aplicación de visualización aparte. Cabe destacar que Flash Paper 2 está mucho más orientado a colaborar con programas como Microsoft Word, Power Point y Excel, todos ellos integrantes de Microsoft office **(WB, 23)**.

Adobe Acrobat 9 Pro

Es una herramienta desarrollada por Adobe Systems para crear, visualizar y manipular archivos PDF **(WB, 24)**. Entre las funciones que nos ofrece se encuentran las siguientes[25]:

- Crear y compartir documentos PDF.
- Combinar archivos de varias aplicaciones.
- Proteger y controlar la información confidencial.
- Eliminar de forma permanente información confidencial.
- Crear formularios y recopilar datos de forma sencilla.
- También permite restringir los permisos de impresión, copia o modificación.

Una vez expuestas las principales características de cada una de las herramientas consideradas, optamos por elegir a Adobe Acrobat 9 Pro como la mejor opción para convertir los documentos del CHEM en documentos PDF. Las razones son las siguientes:

- La documentación que existe es suficiente para poder aprender a utilizar la herramienta a diferencia de la confusa documentación de iText.
- No está ligado a ciertos programas como es el caso de Flash Paper 2 con la paquetería de Office.
- Permite restringir los permisos de impresión y copia.
- Nos brinda más opciones de manipulación de documentos PDF, contrario a Flash Paper 2 o iText las cuales son muy limitadas.

[25] Para más detalle acerca de las características del programa visite el sitio web: http://www.adobe.com/es/products/acrobatpro.

En el capítulo dedicado a la *Seguridad en el CHEM* analizaremos nuevamente cada una de las herramientas pero desde el punto de vista de la seguridad, esto nos permitirá nos sólo elegir la mejor opción para convertir los documentos a documentos PDF sino también la opción que ofrece mayor oportunidades para proteger la información.

5. Generador de estadísticas

Ya decíamos al principio de esta tesis que la primera versión del CHEM contaba únicamente con un generador de concordancias como herramienta de análisis, por lo que para la nueva versión también se acordó agregar una nueva herramienta de explotación de corpus. La herramienta se llama Generador de estadísticas y servirá para medir la asociación que puede existir entre la palabra que se buscó y la que ocurre inmediatamente antes o inmediatamente después. De igual forma se medirá la relación que pueda tener la palabra de búsqueda con todas las palabras que arroja la ventana de la concordancia.

Cabe destacar que la creación de esta nueva herramienta busca también ayudar al proyecto del CHEM a cumplir con una parte de su objetivo principal, la de disponer de herramientas que permitan a investigadores o académicos el análisis del corpus.

En el siguiente apartado describiremos en primer lugar la tabla de la cual se servirá el generador para producir las estadísticas, en segundo lugar en qué consisten cada una de las medidas estadísticas que genera la herramienta y por último algunas de las utilidades que puede tener el generar las medidas mencionadas sobre el corpus.

5.1 Tabla de contingencia

Primeramente para tener una base sobre la cual calcular las medidas estadísticas, es necesario construir una tabla de contingencia. Una tabla de contingencia se utiliza para mostrar la relación que existe entre dos variables.

Tabla 3. Tabla de contingencia para w_1w_2

	w_2	$\overline{w_2}$	total
w_1	$f(w_1w_2)$	$f(w_1\overline{w_2})$	$f(w_1)$
$\overline{w_1}$	$f(\overline{w_1}w_2)$	$f(\overline{w_1w_2})$	$f(\overline{w_1})$
total	$f(w_2)$	$f(\overline{w_2})$	$T = \sum f(w_i)$

En nuestro caso, la tabla de contingencia nos permitirá registrar la relación que puede existir en tres diferentes casos:

1. La palabra que se consultó en el corpus y la palabra que aparece inmediatamente antes.
2. La palabra que se consultó y la palabra que aparece inmediatamente después
3. La palabra que se buscó y cada palabra que pueda aparecer alrededor de ella en la ventana de la concordancia.

Para ayudar a la descripción de la construcción de la tabla de contingencia nos servimos de la Tabla 4 que muestra cómo es que se estructura suponiendo que la palabra que se consultó fue *seguros* y que la palabra que ocurre inmediatamente antes es *tiene*, los valores de la tabla de contingencia surgirían de la siguiente estructura:

Tabla 4. Tabla de contingencia para "seguros".

	$seguros$	$\overline{seguros}$	Total
$tiene$	$f(tiene, seguros)$	$f(tiene, \overline{seguros})$	$f(tiene)$
$\overline{tiene}$	$f(\overline{tiene}, seguros)$	$f(\overline{tiene}, \overline{seguros})$	$f(\overline{tiene})$
Total	$f(seguros)$	$f(\overline{seguros})$	$T = \sum f(w_i)$

La segunda columna contiene la frecuencia de la palabra *seguros*, mientras que la tercera indica las veces que no ocurre *seguros* dentro del corpus (la raya situada sobre la palabra indica que la palabra "no aparece o no ocurre")[26]. El segundo reglón de la segunda columna indica todas las veces que ocurre *tiene* junto con *seguros*, el segundo renglón de la tercera columna indica las veces que *tiene* ocurre sin estar junto con la palabra *seguros*. El tercer renglón de la segunda columna se refiere a las veces que la palabra *seguros* ocurre sin

[26] En otras palabras, la tercera columna contabiliza el número de palabras que no son la palabra *seguros*.

que aparezca la palabra *tiene*, por último el renglón tercero de la tercera columna señala el número de veces que no ocurren juntas ni la palabra *tiene* ni la palabra *seguros*.

En cuanto a los totales, éstos deben coincidir para reflejar que la tabla de contingencia se ha construido correctamente, una vez comprobados, se pueden calcular las medidas estadísticas que describimos a continuación.

5.2 Medidas estadísticas

Para el análisis lingüístico que surgirá de este generador de estadísticas se tomaron en cuenta cuatro medidas principales, esto debido a que son las medidas más populares para medir la asociación en un corpus y a que cada una presenta ventajas y desventajas entre sí. En seguida describimos brevemente estas medidas. Una explicación más amplia de estas medidas se puede encontrar en Medina (2003) y Kageura (1999).

- Prueba de independencia X^2: esta medida es aplicada cuando la tabla de contingencia es de dos por dos, como ejemplificamos en la Tabla 4, y sirve para conocer la asociación que existe entre dos palabras. Esta ecuación calcula para cada celda la variación de los valores esperados bajo la hipótesis de independencia y suma estas variaciones **(Kageura, 1999: 3).** La ecuación para realizar esta prueba es la siguiente:

$$X^2 = \frac{T\left(\left(f(w_1w_2)f(\overline{w_1}\overline{w_2})\right) - \left(f(\overline{w_1}w_2)f(w_1\overline{w_2})\right)\right)^2}{f(w_1)f(\overline{w_1})f(w_2)f(\overline{w_2})}$$

El problema con esta prueba es que no es apropiada para muestras pequeñas o cuando algún valor de la tabla de contingencia es menor a cinco, esto debido a que pierde fidelidad.

- Razón de semejanza: esta medida, al igual que la anterior, sirve para conocer la relación que existe entre dos variables, pero la razón de semejanza a diferencia de X^2 puede ser aplicada sin importar que el tamaño de alguna de las variables sea menor a cinco. En algunos experimentos, se ha comprobado que el resultado de la razón de

semejanza puede ser más fiel cuando el tamaño de las variables es relativamente pequeño[27]. La ecuación que representa esta prueba se muestra a continuación:

$$-2\log\lambda = 2\left[\left(\log\left(L\left(\frac{f(w_1w_2)}{f(w_2)}, f(w_1w_2), f(w_2)\right)\right) + \log\left(L\left(\frac{f(w_1\overline{w_2})}{f(\overline{w_2})}, f(w_1\overline{w_2}), f(\overline{w_2})\right)\right)\right) - \left(\log\left(L\left(\frac{f(w_1)}{T}, f(w_1w_2), f(w_2)\right)\right) + \log\left(L\left(\frac{f(w_1)}{T}, f(w_1\overline{w_2}), f(\overline{w_2})\right)\right)\right)\right],$$

donde

$$L(p,n,k) = n\log(p) + (k-n)\log(1-p)$$

- Coeficiente de coligación de Yule: se basa en una medida estándar de asociación conocida como razón de producto cruzado que se define de la siguiente forma:

$$Y = \frac{\sqrt{a}-1}{\sqrt{a}+1}$$

donde

$$a = \frac{f(w_1w_2)f(\overline{w_1w_2})}{f(w_1\overline{w_2})f(\overline{w_1}w_2)}$$

- Información mutua: el objetivo de esta medida es comparar la probabilidad de que dos eventos ocurran juntos con la probabilidad de que los dos eventos ocurran independientemente **(Kageura, 1999: 4)**. La ecuación es como sigue:

$$I = \log_2\left(\frac{f(w_1w_2)/T}{(f(w_2)/T)(f(w1)/T)}\right)$$

5.3 Aplicaciones de las medidas estadísticas

Una vez dada una breve descripción del generador de estadísticas y expuesta cada una de las medidas estadísticas en los apartados anteriores, resulta importante que el generador de estadísticas mostrará la medida en que están asociadas las palabras, lo que va más allá del

[27] Véase Kyo Kageura (1999: 41) para una explicación más detallada.

simple conteo del número de veces que las palabras aparecen juntas como sucede con el CORDE cuando genera estadísticas. A continuación señalamos algunas de las aplicaciones en las que pueden ser útiles dichas medidas.

- Colocaciones: una colocación es "la ocurrencia de dos o más palabras que se encuentran en un texto y que tienen a ocurrir cercanas en ciertos contextos" **(Medina, Sierra y Garduño 2004: 7).** Las colocaciones parten del resultado que arroja el generador de concordancias y resultan útiles para conocer como se asocian las palabras para formar frases con sentido, lo cual permite construir diccionarios, por ejemplo, especializados.

- Formaciones gramaticales: las medidas también resultan ser útiles cuando se está estudiando una lengua ya que por medio de las asociaciones se pueden identificar construcciones gramaticales, por ejemplo, en un corpus integrado por textos en idioma inglés podría identificarse que un sustantivo regularmente es antecedido por un artículo o que un sustantivo siempre suceda a un adjetivo. Otro ejemplo sería encontrar las construcciones gramaticales que se forman a partir de la palabra *ojo*, como por ejemplo: *taparle el ojo al macho, echarle un ojo*.

- Análisis lexicográfico: al medir el nivel de asociación de las palabras se puede ayudar a dejar más claro el significado de las palabras y esto resulta útil cuando lo que se quiere es construir un diccionario.

- Patrones fraseológicos: las medidas estadísticas permitirán identificar palabras compuestas de segmentos separados y términos multipalabra.

6. Seguridad del CHEM

Al hacer uso de las ventajas que ofrece Internet, como la posibilidad de difundir información, las aplicaciones están expuestas a ataques que pueden ser desde muy sencillos hasta tener consecuencias de gravedad. El caso del CHEM no es la excepción, también es susceptible al robo de información y al mal uso que de ella puedan hacer terceros.

Aunque los ataques no se pueden evitar, sí se pueden disminuir por lo que el objetivo es hacer del CHEM una aplicación tan segura como sea posible para minimizar el impacto de las agresiones a la seguridad. Cualquier ataque representa un peligro a la integridad de la información contenida y requerida por el CHEM, por lo que es importante tener las consideraciones necesarias para evitar la vulnerabilidad de la información.

Para el CHEM contemplamos la adhesión de funciones que permitan el aseguramiento de la información, por ejemplo, la definición de tipos de usuario y los permisos sobre los documentos, descritos en capítulos anteriores. Pero no sólo eso, la seguridad tiene que ver con toda la información del CHEM, desde la forma en la que se recopila hasta la forma como se procesa para ser proporcionada al usuario. En seguida presentamos las formas en las que se trata de asegurar la información.

6.1 Seguridad en el registro de usuarios

La seguridad en el registro de usuarios tiene que ver con la información proporcionada por los usuarios. En este sentido, cuando la información es capturada, el CHEM verifica que el formato sea correcto, que la información sea tan verídica como sea posible y que la misma no signifique un peligro para la integridad de la información que ya está contenida en la base de datos del CHEM; con esto último nos referimos a que la información que se proporciona pueda ser una inyección de SQL.

Una inyección de SQL es una forma de ataque a las aplicaciones, es una transacción que busca filtrarse en los datos para perjudicar la información de la base de datos a partir de introducir instrucciones SQL en algún campo de un formulario que haga conexión con la base de datos. Otra definición es la que nos proporciona Mitnick, la cual señala que una inyección SQL es “un método de ataque que explota un descuido común de programación” **(Mitnick, 2007: 257)**.

Para evitar inyecciones de SQL, es posible introducir algunos procedimientos que rechazan la entrada de cadenas que parezcan peligrosas para la aplicación. Estas funciones pueden denegar, por ejemplo, la entrada al sistema de cadenas que contengan caracteres como comillas, palabras reservadas del propio lenguaje SQL, diagonales, pipes, los símbolos de mayor o menor que, etc.

6.2 Seguridad en los documentos del CHEM

En este apartado hacemos referencia a la información que el CHEM presentará a los usuarios, es decir los documentos. En una sección anterior mencionábamos la importancia que tiene mantener protegidos a cada uno de los documentos del corpus debido a los distintos permisos que el IIUNAM pudiera tener sobre los mismos, por lo que antes de poder mostrar los documentos deberíamos tener en cuenta los siguientes aspectos: los documentos no pueden ofrecer la posibilidad de ser copiados, impresos o guardados por los usuarios.

Como ya indicamos, en el curso de la investigación de esta tesis nos encontramos con diversas herramientas que ofrecían la posibilidad de visualizar documentos en internet de manera que éstos pudieran ser cargados rápidamente manteniendo las referencias y estilos de cada documento XML que conforma el corpus. Sin embargo, con la posibilidad de mostrar los documentos surgen nuevos inconvenientes. Dichos inconvenientes tienen que ver con la seguridad con la que deben contar los documentos para evitar su plagio. Concretamente nos referimos a que los documentos no deberían ser copiados parcial o totalmente, guardados o impresos. En seguida indicaremos algunas características de las tecnologías previamente descritas que tienen que ver con mantener la seguridad de los documentos.

iText

Señalábamos que iText es una API, que permiten generar PDFs dinámicamente. Estas librerías también ofrecen la posibilidad de añadir restricciones a los documentos que son creados. Algunas de las características que ofrece iText para proteger un PDF son **(WB, 20)**:

- Metadatos a los documentos PDF.
- Marcas de agua.

- Protección con contraseña.
- Encriptación.

Además de estas características, también nos encontramos con que esta librería tiene un método para agregar permisos a un documento. Sin embargo, la funcionalidad de este método no está por completo descrita en la documentación de iText por lo que resulta sumamente confuso.

Flash Paper 2

Nos permite convertir casi cualquier documento, en un archivo con SWF o PDF. Entre las posibilidades que incluye para mantener protegidos a los documentos se encuentran:

- Protección con contraseña.
- Posibilidad para restringir los permisos para impresión, copiado, selección y pegado.
- Encriptación.

Adobe Acrobat 9 Pro

Al igual que las herramientas anteriores, Adobe permite la conversión de documentos a un formato PDF. Las características que nos ofrece para mantener la seguridad de los documentos, las mencionamos a continuación:

- Protección con contraseña en los documentos.
- Posibilidad para restringir los permisos de impresión.
- Restricción de permisos para copiado, selección y pegado activando o desactivando dichas opciones. Las restricciones pueden ser posteriormente cambiadas utilizando una contraseña.
- Permisos para que los motores de búsqueda en internet puedan o no acceder a los metadatos del documento.

Ahora, presentamos algunas de las ventajas o desventajas de las herramientas expuestas con el fin de sacar conclusiones que nos permitan elegir la más adecuada.

iText: nos ofrece características como la marca de agua o la encriptación que pudieran ser valiosas para la seguridad de documentos en WEB, pero que sin embargo no son suficientes. También, aunque incluye la posibilidad de restringir los permisos sobre los documentos PDF que genera, estos permisos no resultan viables si requieren de invertir tiempos excesivos en comprender una documentación confusa e insuficiente. Una desventaja sobresaliente es que iText no permite visualizar los documentos si no se tiene instalado el programa de Adobe Reader y esto no siempre puede asegurarse.

Por lo anterior, iText resultaría más útil si los documentos simplemente fueran a ser compartidos entre varios usuarios que quisieran, por ejemplo, identificar algún tipo de propiedad sobre los documentos y no en el caso de ser mostrados en una aplicación en internet sin que se permita la impresión o copia del PDF.

Flash Paper 2: la principal ventaja de esta herramienta es su facilidad de instalación y uso. Cuando se requiere de convertir un archivo, y asignarle permisos o restricciones únicamente se debe recurrir a un par de opciones para solucionarlo.

Otra ventaja se refiere a la documentación que existe para utilizar las opciones que ofrece esta herramienta. En la página de Adobe, se cuenta con la información suficiente para conocer y utilizar Flash Paper 2.

Sin embargo, el primer inconveniente que encontramos con esta herramienta es que está más orientada a convertir documentos realizados con alguno de los diferentes programas de Office, por ejemplo: Word, Excel o Power Point. Cabe destacar que los documentos electrónicos con los que se integra al corpus son documentos XML, por lo esta herramienta no parece capaz de ayudar a solucionar los problemas de seguridad de los documentos.

Esta herramienta es en realidad una impresora virtual por lo que se integra fácilmente a la paquetería de algunos sistemas, y es precisamente en este punto donde surge el segundo inconveniente. Flash Paper 2 no funciona sobre sistemas operativos como por ejemplo: Windows Vista, Windows Seven o cualquier distribución de Linux. Lo anterior significa que si la herramienta llegara a resultar útil, estaríamos atando la conversión de los documentos al sistema operativo Windows XP.

Adobe Acrobat 9 Pro: la primera ventaja de esta herramienta es que brinda, al igual que Flash Paper 2, la posibilidad de agregar restricciones a los documentos PDF de una manera relativamente sencilla. La segunda ventaja es que el Instituto de Ingeniería cuenta con las licencias para la utilización de este software, por lo que el hecho de que esta herramienta necesite ser pagada no representa un problema.

Después de haber descrito y señalado algunas breves, pero precisas, ventajas y desventajas, logramos identificar a Adobe Acrobat como la herramienta que resultaría más útil a nuestro propósito. Como mencionamos en el apartado *Visor de documentos* la herramienta seleccionada para convertir los documentos XML del corpus a documentos con formato PDF fue por factores como la facilidad de instalación, facilidad de uso, compatibilidad y disposición. Ahora una vez analizadas las ventajas que en cuanto a seguridad nos ofrece cada una de las herramientas analizadas reafirmamos la decisión de usar Adobe Acrobat 9 Pro debido a que resulta ser la mejor alternativa para mantener protegidos los documentos del CHEM.

7. Ajax

Con el desarrollo de esta tesis, será necesario reacomodar los contenidos de la interfaz gráfica y dar cabida a las nuevas herramientas de análisis y opciones de gestión del CHEM. En el planteamiento del problema de esta investigación decíamos que la primera versión del CHEM requería de contenidos que proporcionaran mayor información del proyecto, el desarrollo de nuevas herramientas, así como la creación de opciones para gestionar a las mismas. Una vez realizadas cada una de las éstas tareas será necesario dotar de eficacia a la interfaz, para esto proponemos la adhesión de un conjunto de tecnologías denominadas AJAX.

7.1 Definición

AJAX es el acrónimo utilizado para *Asynchronous JavaScript And XML*, lo cual significa Java Script asíncrono y XML. Básicamente AJAX hace uso del objeto XMLHttpRequest basado en Java Script para responder a peticiones del servidor Web de manera asíncrona sin tener que recargar toda la página. Haciendo uso de XMLHttpRequest, las aplicaciones Web pueden recolectar o enviar información al servidor. El servidor la procesa y manipula dependiendo de las necesidades y luego cambia el aspecto de la página Web dinámicamente sin que se tenga que refrescar por completo **(Babin, 2007)**.

Las aplicaciones que utilizan AJAX se ejecutan en el navegador del cliente pero manteniendo la comunicación asíncrona con el servidor, es decir, los datos que el usuario requiere se solicitan al servidor y se ejecutan en segundo plano sin interferir con la vista que ya se tiene de la página. AJAX permite entonces que las aplicaciones tengan mayor velocidad agregando también mayor interactividad al usuario.

Para entender el uso de esta técnica ejemplificamos con el siguiente caso:

> *Supongamos que un usuario ingresa a una página, que no utiliza AJAX, primero la página es cargada, después el usuario proporciona sus datos a un formulario o da clic en un enlace. La acción del usuario es enviada al servidor para ser procesada, el usuario espera a que le sea enviado un resultado que finalmente será presentado recargando toda la página*.

Lo que pasaría con AJAX sería lo siguiente:

El usuario entra en una página, la página es cargada, luego el usuario proporciona información a un formulario o da clic en un enlace. La acción es enviada y el usuario espera su respuesta y finalmente la página recarga únicamente en la parte que debería mostrar el resultado de la acción.

Es en el segundo ejemplo donde el conjunto de tecnologías AJAX agrega rapidez a la capacidad de respuesta de un sistema ya que se encarga de recargar únicamente lo que el usuario solicita al servidor. En el capítulo de esta tesis dedicado a la integración de la aplicación se detallará el funcionamiento y la forma en que estas tecnologías son utilizadas.

8. El libro de visitas

Una pregunta que se ha venido haciendo el GIL desde la liberación de la primera versión del CHEM tiene que ver con ¿quién usa el sistema? pero, ¿para qué saber esto? La respuesta es sencilla, saber quién usa el sistema permitiría al GIL ofrecer nuevas posibilidades a los usuarios tomando en cuenta datos como su país o su institución de procedencia. Por ejemplo, si se conociera que el CHEM está siendo visitado en gran medida por usuarios angloparlantes, el sistema podría ser dotado de una interfaz en idioma inglés.

Para contestar a la pregunta ¿quién usa el sistema? Se decidió desarrollar en ese entonces un libro de visitas, el cual es en realidad una lista de registros compuestos por los accesos de aquellos usuarios que han ingresado en las páginas del CHEM[28].

El libro se integra por datos como nombre, correo, institución y número de accesos de cada usuario, por lo que se muestra como la Figura 8.

Nombre	Correo	País	Institución	Ocupación	Accesos
CarlosMéndez	cmendezc@iingen.unam.mx	mx	UNAM	Estudiante	85
ClaudiaPatino Agreda		mx		Estudiante	55
AlfonsoMedina	amedinau@iingen.unam.mx	mx	Instituto de Ingeniería	Investigador	52
DorienNieuwenhuijsen	Dorien.Nieuwenhuijsen@let.uu.nl	nl	Universidad de Utrecht	Profesor	17
GerardoSierra	gsierram@ii.unam.mx	mx	Instituto de Ingeniería, UNAM	Investigador	16
ManuelGómez	alejandroneet@gmail.com	mx		Estudiante	10
EsperanzaMartínez Herrera	dbzmonse@yahoo.com.mx	mx	UNAM	Estudiante	9
alejandronavarro gonzalez	alng_1988@hotmail.com	mx	UAM-Iz	Estudiante	7
JagodaKowalska	jagienka1983@wp.pl	pl	Universidad de Poznań	Estudiante	6
MiriamBouzouita	miriam.bouzouita@kcl.ac.uk	uk	Kings College London	Estudiante	6
Teresita AdrianaReyes Careaga	cuariqui_3@yahoo.com.mx	mx	UNAM	Estudiante	6
Ariadna CarolinaHernández Angulo	andaira617@yahoo.com.mx	mx	Facultad de Filosofía y Letras	Estudiante	5
HeidiDueñas Bastida	heidi_d_b_@hotmail.com	mx	UNAM	Estudiante	5
AlejandraChvarria Amezcua	alejandra3101@hotmail.com	mx	UNAM	Estudiante	4
ELISABETHFERNÁNDEZ MARTÍN	elisabethfm@ugr.es	es	UNIVERSIDAD DE GRANADA	Investigador	4

Figura 8. Libro de visitas del CHEM (primera versión)

Sin embargo, el problema con el libro de visitas de la primera versión es precisamente, como ya mencionamos, que registra los accesos de los usuarios que ingresan al CHEM, lo cual no deja saber si realmente el corpus fue consultado, ya que un usuario pudo haber ingresado y no realizar alguna búsqueda con el generador de concordancias. De poco serviría al GIL saber el número de veces que un usuario accede al CHEM sin saber cuando en verdad lo usa.

[28] Destacamos que al decir "usuarios que han ingresado en las páginas del CHEM" nos referimos a que el usuario previamente ha concluido su registro y ahora puede proporcionar un correo electrónico y una contraseña para acceder al sistema.

Derivado de lo anterior, uno de los propósitos de esta tesis es mejorar el libro de visitas existente contabilizando únicamente cuando los usuarios realizan alguna consulta, es decir una vez que hayan utilizado el generador de concordancias. Destacamos que los usuarios pueden estar o no registrados debido a que la herramienta generador de concordancias está disponible, aunque con reservas, también para usuarios no registrados. El libro de visitas por lo tanto cambia su nombre, después del desarrollo de esta tesis, para llamarse en adelante Registro de consultas.

Éste registro de consultas se dividirá en dos grupos con el objetivo de facilitar su análisis y mejorar la presentación de los mismos. Los grupos los indicamos a continuación:

- Registro de consultas de usuarios registrados.
- Registro de consultas de usuarios anónimos.

En el capítulo de integración de esta tesis expondremos la diferencia que existe entre cada uno de los grupos así como la forma en que se estructuran.

9. Integración final

Ya que hemos analizado cada problema de la primera versión del CHEM, así como investigado y propuesto algunas soluciones a los mismos, nos damos a la tarea de integrar cada una de ellas en la aplicación. Así, en este capítulo nos enfocaremos en describir la forma en que finalmente cada herramienta se vuelve parte del CHEM para crear la nueva versión del sistema.

9.1 Administración de usuarios

En esta sección, indicamos como se construyó e integró la administración de los usuarios del CHEM, comenzando con las mejoras al registro de usuarios y continuando con la función para crear nuevos tipos de usuarios en la aplicación.

9.1.1 Registro de usuarios

Como previamente nos habíamos referido al registro de usuarios y a sus deficiencias, en este apartado expondremos como es que aquellos defectos se corrigieron para optimizar la nueva versión del CHEM.

En primer lugar, el registro de usuarios ahora cuenta con diversas funciones para validar todos los tipos de datos que son ingresados en la base de datos, de esta forma se intenta evita procesar información que pudiera ser peligrosa o falsa y por lo tanto inútil para el administrador del CHEM. Las distintas formas en que un campo puede ser validado pueden consistir en una o varias de las siguientes[29]:

- Validar solo letras en un campo, por ejemplo para el caso de nombres y apellidos.
- Validar sólo números en un campo, por ejemplo para los campos de fecha.
- Validar formato de correo electrónico, es decir que el correo se componga del nombre de usuario, el símbolo "@", el cliente de correo electrónico, punto (.) y por último el dominio.

[29]Es relevante mencionar que todas las formas en las que pueda ser ingresada información al CHEM mediante formularios en las interfaces son validadas con las mismas funciones que en principio se construyeron para el registro de usuarios.

- Validar campos vacios, esta función se aplica para que todos los datos que son requeridos al usuario sean proporcionados por el mismo.

- Validar fechas, en este caso se valida que la fecha que el usuario proporciona exista en el calendario, por ejemplo en el caso de años bisiestos.

Los datos son validados de dos formas: primero haciendo uso de funciones Java Script y luego con funciones creadas en el lenguaje Java.[30] Las funciones creadas con Java Script nos permiten verificar el formato de los datos e indicar al usuario cuando éstos son correctos o incorrectos. Las funciones Java se encargan, además de validar el formato, de validar que las cadenas no contengan estructuras propias del lenguaje SQL, por ejemplo: “Delete “ o “ Insert ”, de esta forma se busca evitar a las llamadas inyecciones SQL.

La forma en la que los usuarios proporcionan sus datos para registrarse es mostrada a continuación:

Figura 9. Formulario para registro de usuarios

[30] Para justificar el uso de la validación tanto a nivel cliente (Java Script) como a nivel servidor (Java) pensamos, por ejemplo, en que un usuario malintencionado pudiera desactivar Java Script de su navegador y de esta forma enviar datos erróneos a la base de datos.

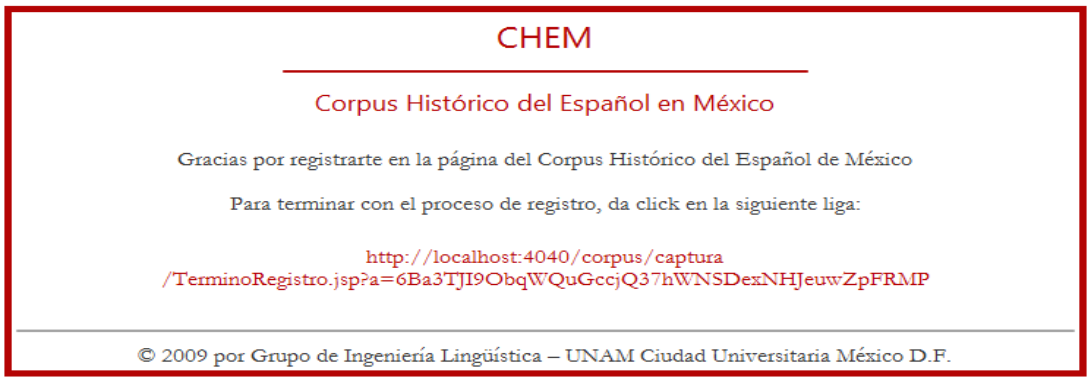
Una vez proporcionados los datos con el formato correcto, es enviado un correo electrónico al usuario para validar sus datos y activar su cuenta como ya dijimos, esto asegura que el registro fue realizado por una persona. Para concluir con el tema del registro de usuarios, destacamos que por *default* todos los usuarios son en principio registrados como usuarios de tipo "básico".

9.1.2 JavaMail para confirmación de registro y recuperación de contraseña

Anteriormente describimos y justificamos la utilización de la herramienta JavaMail así como presentamos brevemente la forma en la que ésta sería utilizada para la nueva versión del CHEM. A continuación nos enfocaremos a definir la forma en que JavaMail será integrada al CHEM tanto para el envío de un correo electrónico que permita la activación de la cuenta del usuario, como para el envío de un correo con la contraseña de un usuario si este la solicita. Por último, detallaremos cómo funciona esta API.

El proceso de confirmación y activación de cuentas inicia una vez que el usuario acepta sus datos. Los datos del usuario son guardados en la base de datos, añadiéndose a éstos un número de registro aleatorio de 40 caracteres y un valor por defecto "desactivado" al estado de usuario. El estado de usuario, así como el número de registro, determinan a los usuarios que pueden hacer uso o no del CHEM.

Si el proceso de registro es correcto, se envía un correo electrónico al usuario como muestra la Figura 10. El correo incluye el número de registro de usuario, el cual se presentará como un enlace que permitirá terminar con su registro. Una vez que el usuario de clic en el enlace, la base de datos actualizará el registro del usuario, el valor del número de registro se anula y el estado del usuario cambia su valor a "activo" para ahora permitir iniciar una sesión y hacer uso del CHEM con las ventajas de ser un usuario registrado.


CHEM

Corpus Histórico del Español en México

Gracias por registrarte en la página del Corpus Histórico del Español de México

Para terminar con el proceso de registro, da click en la siguiente liga:

http://localhost:4040/corpus/captura
/TerminoRegistro.jsp?a=6Ba3TJI9ObqWQuGccjQ37hWNSDexNHJeuwZpFRMP

© 2009 por Grupo de Ingeniería Lingüística – UNAM Ciudad Universitaria México D.F.

Figura 10. Correo de activación de cuenta

También puede ser enviado un correo electrónico al usuario si éste olvida su contraseña. En este caso, el CHEM permite recuperar la contraseña del usuario de la siguiente forma. Una vez que el usuario proporciona su fecha de nacimiento y correo electrónico con el que se registró, le es enviado un mensaje a su correo conteniendo la contraseña con la que hizo su registro. La Figura ilustra la forma en la que el usuario recibe el mensaje de correo con su contraseña.

Figura 11. Correo para recuperar contraseña

Una vez explicado lo anterior, en seguida indicamos brevemente como se construye la clase para el envío de correo electrónico utilizando la API JavaMail. En primer lugar se debe indicar lo siguiente: **(WB, 25)**:

- El nombre del servidor de correo electrónico.
- TLS[31], si está disponible, es decir si se requiere conexión segura.
- Puerto del servidor que permite el envío de correo electrónico.
- Nombre del usuario propietario de la cuenta que enviará el correo: el remitente.

[31] Seguridad de la capa de transporte o Transport Layer Security (TLS, por sus siglas en inglés) es un protocolo que proporciona comunicaciones seguras por redes como Internet.

- Señalar si se requiere que haya autenticación, es decir, contraseña y usuario para conectarse.

Una vez realizado lo anterior, podemos empezar a crear el mensaje estructurándolo de la siguiente forma:

- Agregamos el asunto del correo electrónico.
- Indicamos quien envía el correo: dirección y nombre del usuario.
- Definimos el tipo de contenido que será mostrado en el correo (HTML) así como el texto o imágenes del correo.

Por último, para terminar de construir la clase enviamos el mensaje como sigue:

- Señalamos el nombre del protocolo que utilizamos para enviar el correo, en otras palabras, el protocolo SMTP[32].
- Establecemos la conexión de acuerdo al nombre de usuario y contraseña para poder enviar el mensaje.
- Finalmente cerramos la conexión.

Para utilizar JavaMail se programó un método de envió de correo electrónico que recibe dos parámetros: la dirección de correo del usuario (destinatario) y el número aleatorio de registro.

9.1.3 Cambio de tipo de usuario

Señalábamos que una vez que el usuario cumple con su registro, el sistema por default lo considera como un usuario de tipo “básico”, pero esto puede ser cambiado, posteriormente, únicamente por el administrador del CHEM haciendo uso de una interfaz construida especialmente para este propósito.

[32] El protocolo simple de transferencia de correo (SMTP, por sus siglas en inglés), permite el intercambio de mensajes de correo electrónico entre computadoras.

La interfaz para modificar el tipo de usuario permite proporcionar un patrón de búsqueda para recuperar todos los usuarios que coinciden con él, el patrón puede contener fragmentos de los datos del usuario por ejemplo: nombres o parte del correo electrónico, la Figura 12 ilustra un caso de búsqueda en esta interfaz.

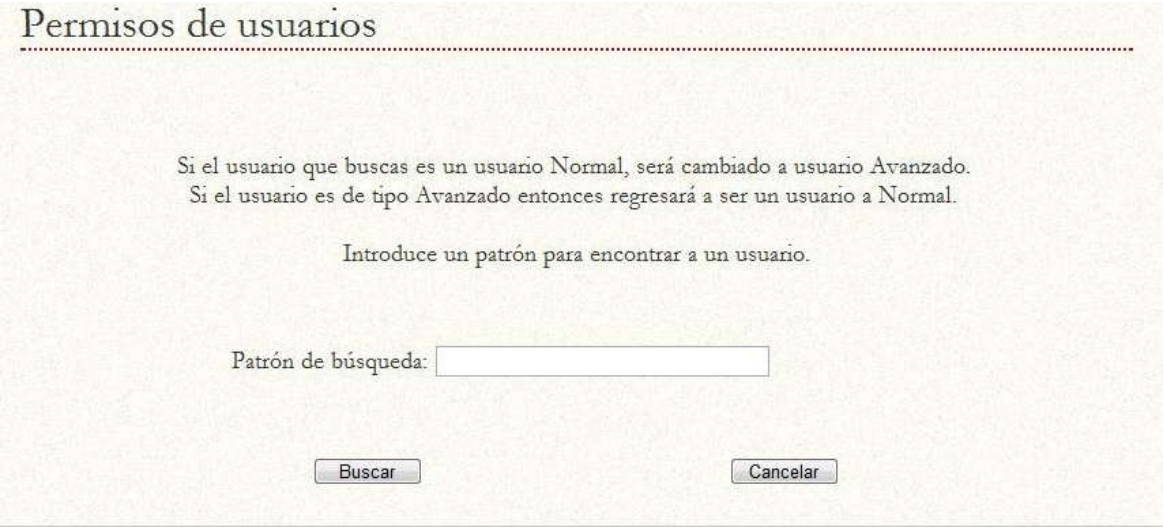


Figura 12. Interfaz de búsqueda de usuario por patrón

Una vez realizada la búsqueda, la interfaz muestra a todos los usuarios que contengan algún dato que concuerde con el patrón de búsqueda (la Figura 13 ejemplifica lo dicho) y el administrador selecciona aquel a quien desee cambiar el tipo de usuario. Aclaramos que estrictamente, si el usuario seleccionado es un usuario básico, su tipo de usuario cambia a avanzado y viceversa.

Figura 13. Interfaz de usuarios que concuerdan con el patrón de búsqueda

9.1.4 Creación de tipos de usuario

Con el objetivo de lograr una administración lo más óptima posible para el CHEM, en la nueva versión se decidió tipificar a los usuarios del mismo, quedando divididos en los tipos de usuarios: Administrador, Básico, Avanzado, Anónimo y Anónimo UNAM. Sin embargo, contemplando el crecimiento del CHEM a corto y largo plazo también se desarrolló una interfaz que permite la creación de nuevos tipos de usuarios.

La interfaz para la creación de tipos de usuarios permite, mediante un formulario, ingresar el nombre del nuevo tipo de usuario. Una vez hecho esto, es posible asignar permisos a este nuevo usuario sobre el uso de herramientas y documentos del corpus.[33] Subrayamos que ésta interfaz (Figura 14) únicamente puede ser utilizada por el usuario administrador del CHEM.

[33] Más adelante describiremos cómo es que se realiza la asignación de permisos de herramientas y documentos a los usuarios en la nueva versión del CHEM.

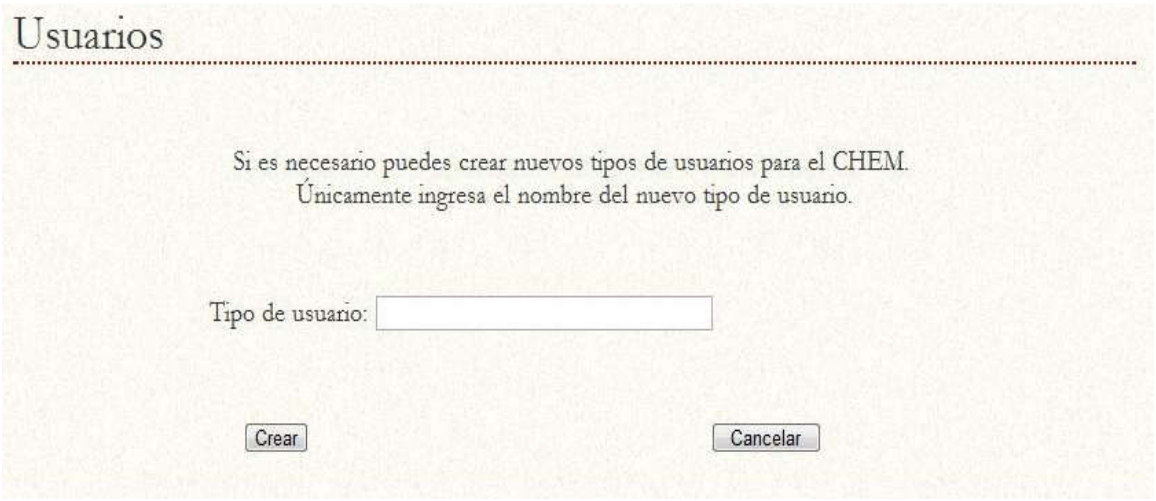


Figura 14. Interfaz para crear nuevos tipos de usuario

9.1.5 Método para la obtención de la dirección IP

Para lograr obtener la dirección IP de los usuarios finalmente se optó por el método desarrollado por la página Geobytes.com, debido a que es un método probado que además, nos brinda la información que se requiere para cumplir con el objetivo de distinguir a los usuarios del CHEM dentro de la UNAM.

La forma en que funciona este método consiste en lo siguiente:

Un usuario accede al CHEM e inmediatamente un método de la página Geobytes.com captura su dirección IP y el nombre del país desde el cual el usuario está consultado la aplicación.

Una vez hecho lo anterior, la aplicación procesa los datos provistos por Geobytes y los discrimina para saber si el usuario esta accediendo a la aplicación dentro o fuera de la UNAM[34]. Por último la aplicación concede algunos permisos sobre los materiales del CHEM si es que el usuario se encuentra dentro de la UNAM.

[34] Para saber si la dirección IP del usuario se encuentra dentro de la UNAM, se verifica si la IP cumple con determinada condición

9.2 Administración de documentos

En secciones anteriores decíamos que para la nueva versión del CHEM se contaría con tipos de permisos asignados a los documentos, esto con el objetivo de mantener en la medida de lo posible la seguridad de los mismos, limitando la posibilidad de que éstos puedan ser vistos por todos los usuarios.

Sin embargo, teniendo en mente la idea de que el CHEM seguirá aumentando el número de documentos que lo integran así como las restricciones a los mismos, se contempló, para la nueva versión, una interfaz que permite la creación de nuevos tipos de permisos.

9.2.1 Creación de permisos de documentos

Los permisos que se contemplaron para la nueva versión del CHEM son los siguientes:

- Documentos con permiso
- Documentos sin permiso
- Documentos con permiso en trámite
- Documentos que no requieren permiso
- Documentos con permisos parciales

Sin embargo, decíamos que el CHEM seguirá aumentando su tamaño, así que podría ser necesaria la creación de nuevos tipos de permisos. Para resolver este problema, la nueva versión contiene una interfaz para crear tipos de permisos de documentos. Esta interfaz, aunque sencilla, repercute en temas tan importantes como la asignación de estos permisos a los usuarios, que posteriormente describiremos. Básicamente la interfaz contiene un cuadro de texto donde el usuario proporciona el nombre del nuevo tipo, una vez hecho esto la base de datos es afectada y en adelante se reconoce el nuevo permiso de documentos. La Figura 15 ilustra la interfaz.

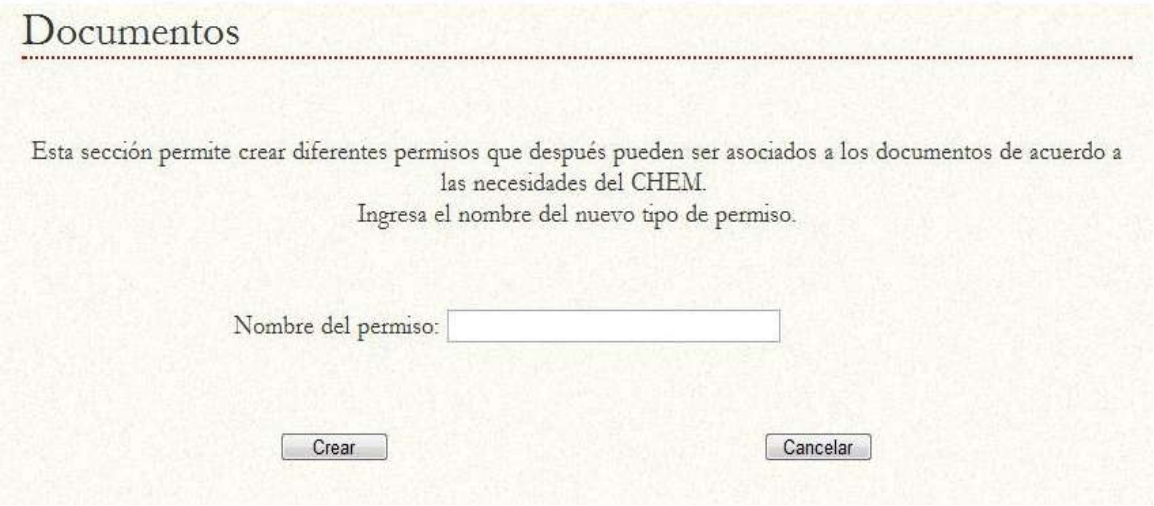


Figura 15. Interfaz para crear permisos de documentos

9.2.2 Asignación de permisos de documentos a usuarios

En secciones anteriores indicábamos los tipos de usuarios y permisos de documentos que se contemplan para la nueva versión del CHEM; de igual forma hemos descrito en esta investigación la manera en que podrían ser agregados tanto nuevos tipos de usuarios como permisos de documentos, considerando el crecimiento del CHEM a largo plazo. Es momento de exponer el procedimiento en que los permisos de documentos se pueden relacionar con los tipos de usuarios para conceder o negar privilegios a estos últimos.

Con el objetivo de permitir la relación antes mencionada (asignar permisos de documentos a los tipos de usuario), se han creado un par de interfaces que funcionan del siguiente modo: la primera muestra al usuario administrador una lista con todos los tipos de usuarios que existen en la aplicación hasta ese momento (usuarios anónimos, usuarios avanzados, etc.) como muestra la Figura 16.

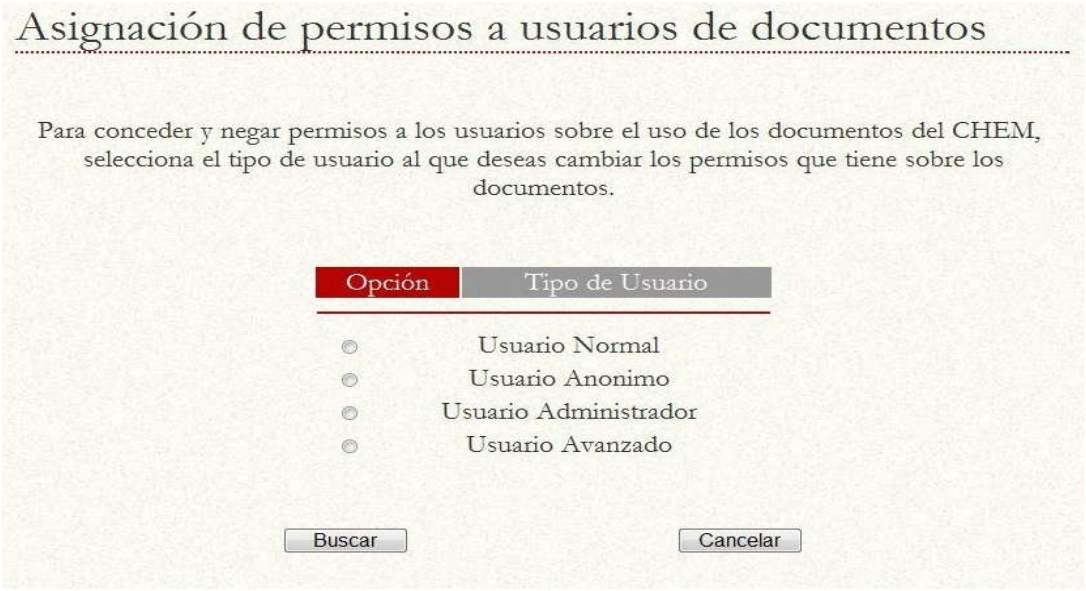


Figura 16. Asignación de permisos a usuarios de documentos

Una vez que el administrador elige al usuario al cual desea cambiar los permisos surge una segunda interfaz, la cual contiene los diferentes permisos de documentos que pueden tener los usuarios así como una casilla para activarlos o desactivarlos según se requiera. (

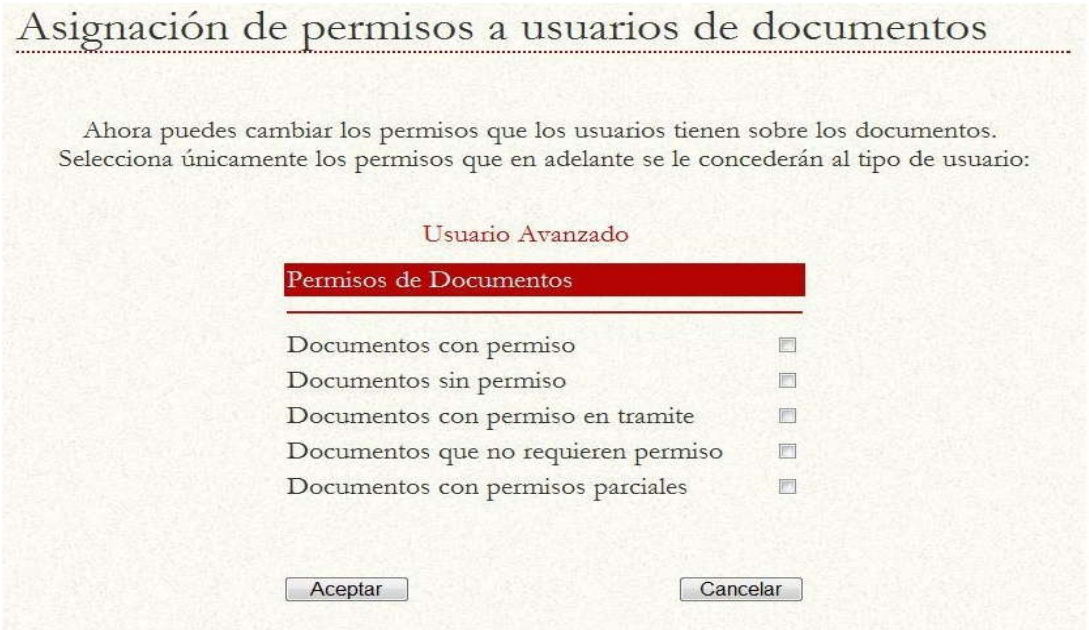


Figura 17. Usuarios existentes en la BD para cambio de permisos de documentos

9.3 Administración de herramientas

9.3.1 Creación de herramientas

Recordamos que las herramientas con las que contará la nueva versión CHEM destinadas para ser utilizadas por los usuarios, registrados o no registrados, son las siguientes: visor de documentos y generador de concordancias. Sin embargo, al igual que para la creación de permisos de documentos, tema previamente descrito, el CHEM ahora cuenta con una interfaz que en este caso permite la adhesión de nuevas herramientas a la aplicación para que posteriormente estas también puedan ser administradas.

La justificación para la creación de la interfaz para agregar herramientas al CHEM es que sabemos que el CHEM seguirá creciendo de forma indefinida una vez concluida ésta investigación tanto en herramientas como en contenidos.

La manera como trabaja la interfaz es similar al funcionamiento de la que se utiliza para agregar nuevos permisos de documentos, es decir, ésta incluye un cuadro de texto en el cual el usuario administrador puede proporcionar el nombre de la nueva herramienta de explotación del corpus como muestra la Figura 18.

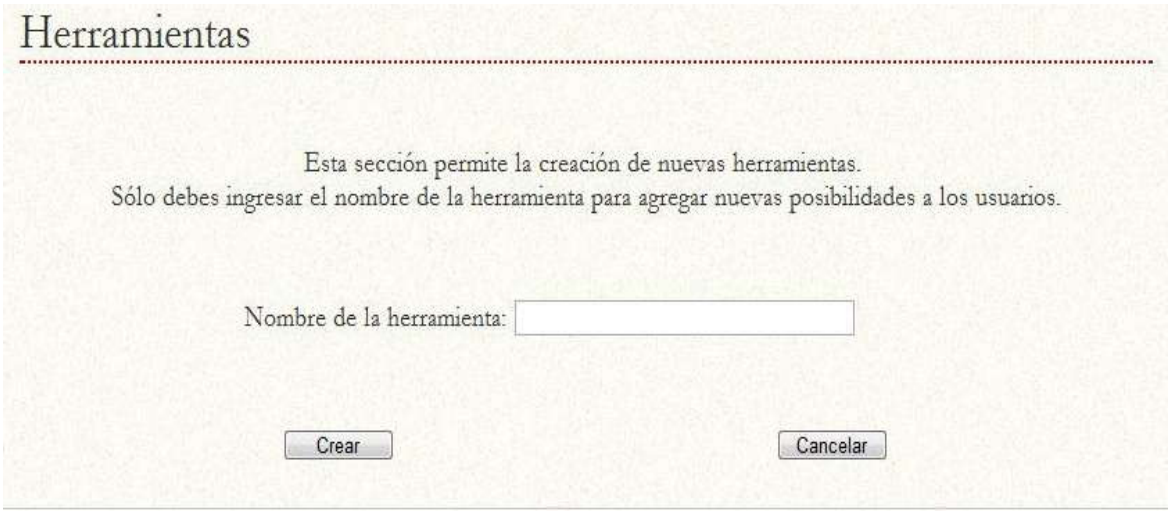


Figura 18. Interfaz para crear herramientas

9.3.2 Asignación de permisos de uso de herramientas a tipos de usuario

Una de las razones por las cuales se contempló la tipificación de usuarios para la nueva versión del CHEM fue la de limitar el uso de los documentos y herramientas que lo conforman con el fin de administrar y mantener cierto nivel de seguridad para los mismos.

Continuando con la administración del CHEM, señalamos que las herramientas, al igual que los documentos, requieren ser asignadas por el administrador a los tipos de usuario con la finalidad de que éstos puedan tener acceso a las mismas. El administrador, precisamos, asigna los permisos de uso de acuerdo a la conveniencia del proyecto del CHEM. En seguida describimos la forma en que serán asignados dichos permisos a los diferentes tipos de usuario.

El procedimiento es similar al que se realiza cuando se quiere asignar permisos de documentos a los tipos de usuario, es decir, fueron desarrolladas una serie de interfaces que permiten la administración de los permisos de las herramientas contra los usuarios, donde la primera interfaz muestra a todos los tipos de usuario registrados en el corpus y la opción de elegir a alguno de ellos para modificar sus permisos. Una vez que el usuario administrador selecciona a algún usuario, una segunda interfaz muestra una lista de las herramientas del corpus así como casillas que indican, en caso de estar seleccionadas, que el usuario tiene permitido utilizar esa herramienta. Por otra parte, si las casillas no están seleccionadas indican lo contrario, o sea que el uso de esa herramienta no está permitido para ese usuario. La Figura 19 ilustra y nos ayuda a comprender como es que se presenta esta segunda interfaz al usuario.

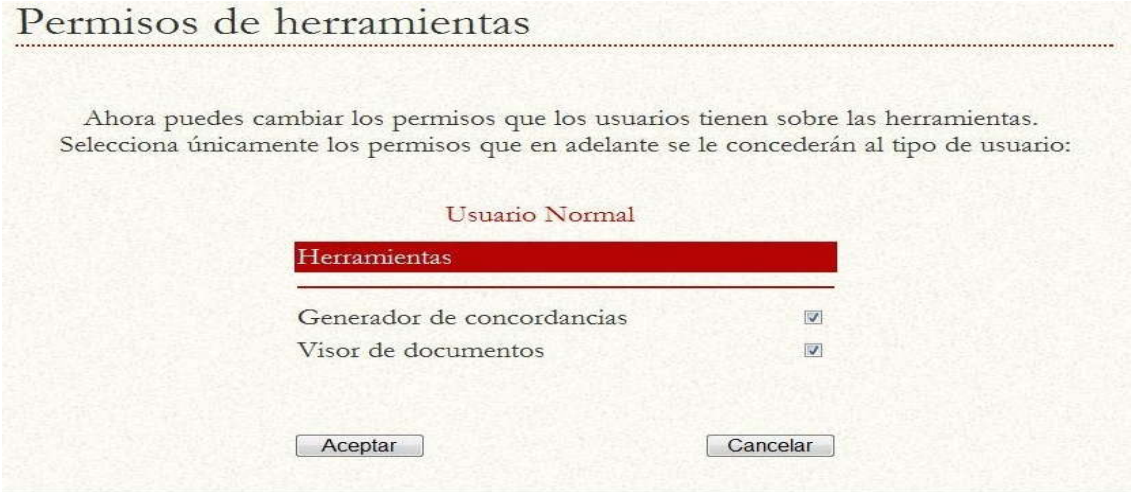


Figura 19. Asignación de permisos de herramientas a un usuario

9.4 Generador de estadísticas

Para describir la forma en que fue integrado el generador de estadísticas al CHEM recurriremos primero a describir cómo es que son generadas las estadísticas una vez creada la concordancia. Posteriormente nos serviremos de algunas imágenes para ilustrar la integración.

En primer lugar, para poder generar las estadísticas se creó una clase Java en la cual se programaron los métodos necesarios para llevar a cabo el procesamiento de la concordancia que derivará en las medidas estadísticas. El primer método de la clase, recibe tres parámetros, mismos que describiremos a continuación:

1. Palabra buscada: se refiere a la palabra que consulta el usuario para generar las concordancias.

2. Forma de la palabra buscada: por lo general las palabras en el corpus están categorizadas en ciertas formas de palabras, por ejemplo: lema (*l*), palabra en ortografía normalizada (*n*), palabra ortográfica (*o*) y trascripción fonológica (*f*). Para la generación de estadísticas únicamente se procesan aquellas palabras de la concordancia que son del tipo *n*.[35]

3. Arreglo bidimensional: este arreglo tiene en sus posiciones a cada palabra que acompaña a la concordancia. En este arreglo destacan dos características: la primera, que la palabra pivote, es decir, la palabra que se consultó y que generó la concordancia, se localiza entre dos pico paréntesis dentro del arreglo (<>); y la segunda, que cada una de las filas del arreglo corresponde a cada una de las filas de la concordancia.

Con lo anterior, el método asigna a cada una de las palabras dentro de tres grupos: 1) palabras inmediatamente antes de la palabra pivote, 2) palabras inmediatamente después de la palabra pivote y, 3) palabras que se encuentran en toda la concordancia. Una vez que las

[35] Aunque el generador de estadísticas procesará, después de esta investigación, únicamente a aquellas palabras que sean del tipo *n*, se contempla que en un futuro esa condición pueda cambiar, por lo que podrían procesarse también aquellas que sean del tipo *o*, *f* ó *l*.

palabras están dentro de los grupos se llama a un nuevo método que recibe los siguientes parámetros:

1. Uno de los grupos de palabras: nos referimos a alguno de los grupos de palabras que construye el método anterior.

2. El tamaño de la concordancia: es decir, el número de renglones que integran la concordancia.

3. La forma de palabra: como mencionamos antes, el generador de estadísticas únicamente procesará aquellas palabras de la forma *n*.

El método se ejecuta tres veces, una por cada grupo de palabras y básicamente busca la frecuencia de todo el corpus relacionada con la forma de palabra que procesa el generador. Luego el método busca de cada palabra del grupo su frecuencia en la base de datos, tomando en cuenta únicamente a aquellas palabras que tienen el tipo n.

A continuación, el método se encarga de realizar las operaciones para elaborar las medidas estadísticas: información mutua (I), razón de semejanza ($-2\log\lambda$), prueba de independencia (X^2) y prueba de coligación de Yule (Y). Luego, el generador se encarga de producir un promedio normalizado ordenado en forma descendente para cada resultado. Para generar el promedio normalizado se obtiene el resultado de mayor valor de cada una de las medidas estadísticas, después, todos los resultados son divididos entre el resultado de mayor valor, por último se suman todos los resultados generados de las divisiones para luego ser divididos entre el número de medidas estadísticas que se realizaron, en este caso cuatro.

En caso de que alguna de las medidas estadísticas no se pudiera realizar, se contempló agregar las abreviaturas N/S (No significativo) y N/C (No calculable). La primera es usada para los casos en los que se infrinja alguna regla matemática o de la propia fórmula, por ejemplo, la división entre 0 ó el cálculo de la prueba de independencia cuando los valores son menores a 5. Por su parte, N/C se muestra cuando no es posible realizar ninguna de las medidas estadísticas, lo cual significa que no se puede realizar un promedio. Finalmente, se genera una tabla para cada grupo de palabras con las medidas estadísticas, las siguientes figuras ilustran cada una de las tablas resultado del procesamiento del generador de concordancias.

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
tiene	2.4818	3.5551	N/S	0.6307	1
bjvan	2.1941	2.9885	N/S	0.5787	0.8808
esten	1.6833	2.0366	N/S	0.4757	0.6684

N/S = No Significativo
N/C = No Calculable

Figura 20. Estadísticas de palabras inmediatamente antes

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
sino	2.7331	4.067	N/S	0.673	1
escrjviendo	1.5435	1.7902	N/S	0.4448	0.5553
con	1.4602	1.6468	N/S	0.4258	0.524

N/S = No Significativo
N/C = No Calculable

Figura 21. Estadísticas de palabras inmediatamente después

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
los	3.2958	11.5592	N/S	0.8222	0.9735
oy	3.5804	5.939	N/S	0.7983	0.8283
no	2.6672	8.8212	N/S	0.7463	0.8053
que	2.6027	8.5513	N/S	0.7376	0.788
su	2.1564	6.7289	N/S	0.6716	0.6671
sino	2.7331	4.067	N/S	0.673	0.6446
y	1.8788	5.632	N/S	0.6244	0.5905
tiene	2.4818	3.5551	N/S	0.6307	0.5893
senorios	2.3765	3.3452	N/S	0.6122	0.5659
bjvan	2.1941	2.9885	N/S	0.5787	0.5251
reinos	2.1141	2.8346	N/S	0.5635	0.507
onbres	2.04	2.6936	N/S	0.5492	0.4902
amparo	1.7887	2.2266	N/S	0.4982	0.4327
les	1.7887	2.2266	N/S	0.4982	0.4327
en	1.6833	2.0366	N/S	0.4757	0.4083
esten	1.6833	2.0366	N/S	0.4757	0.4083
escrjviendo	1.5435	1.7902	N/S	0.4448	0.3757
con	1.4602	1.6468	N/S	0.4258	0.3561
casa	1.2778	1.3432	N/S	0.3827	0.3129

N/S = No Significativo
N/C = No Calculable

Figura 22. Estadísticas de palabras dentro de toda la concordancia.

9.5 Visor de Documentos

Para continuar con el capítulo de integración final de esta tesis, resultado de la investigación en capítulos anteriores, explicamos ahora cómo es que funciona el visor de documentos que formará parte de la nueva versión del CHEM.

Anteriormente habíamos resuelto utilizar la herramienta Adobe Acrobat Professional para convertir los documentos del corpus en documentos tipo PDF, esto por múltiples ventajas, entre ellas la seguridad, la disposición y la facilidad de uso.

El visor de documentos es una herramienta que permite que el usuario pueda visualizar los documentos completos del CHEM, siempre y cuando dicho usuario tenga permiso para verlos o que los documentos no requieran permiso para ser expuestos. El visor funciona en el CHEM como a continuación describimos:

El CHEM tiene ahora una nueva opción llamada “Acervo”, ésta opción permite que el usuario pueda ver una lista de los documentos que forman parte del corpus y, como mencionamos, siempre y cuando pueda verlos, la Figura 23 ilustra una lista de documentos como se muestra al usuario. Una vez que el usuario selecciona el título de alguno de los documentos de la lista, un servlet se encarga de mostrar el documento dentro de un espacio en la página web como muestra la Figura 24.

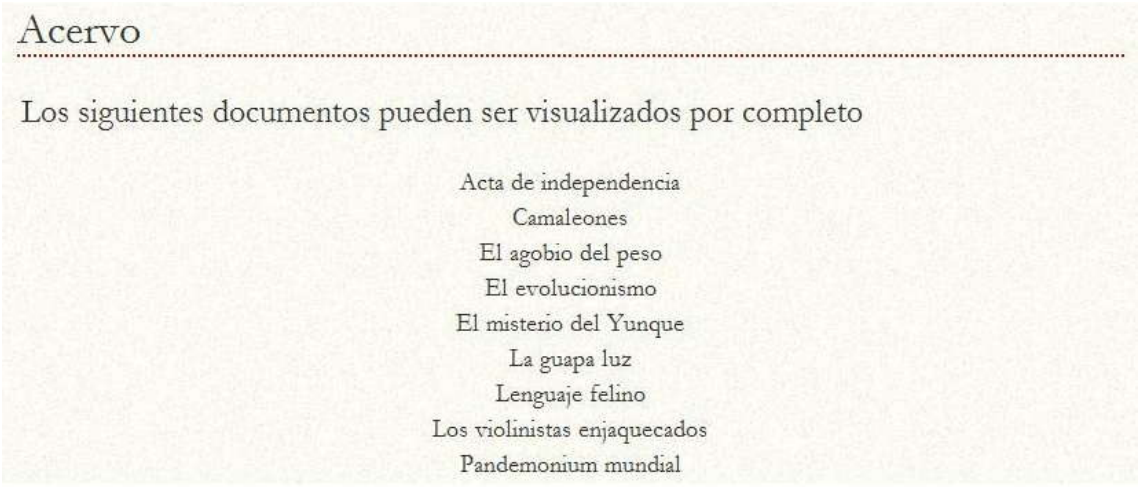


Figura 23. Ejemplo de lista de documentos disponibles para un tipo de usuario

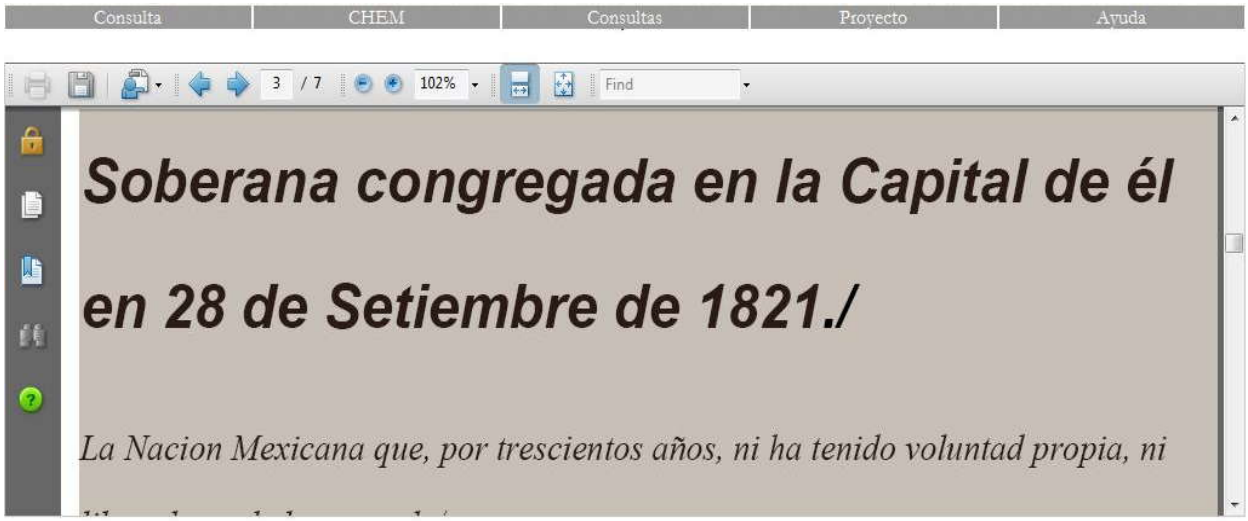


Figura 24. Documento incrustado en la página del CHEM

El servlet que se encarga de mostrar los documentos contiene la instancia de una clase Documento, esta instancia permite contar el número de veces que es visto. En la instancia se indican el *content type* y las cabeceras del documento, posteriormente se concatena la ubicación de los documentos al nombre del documento; este nombre se recibe de la lista que se mostró previamente. Después se pide el tamaño del archivo PDF que se va a mostrar y el servlet se encarga de imprimir el documento hasta que este es cargado en su totalidad. Asimismo indicamos que el servlet se encarga de mostrar únicamente el documento si el usuario tiene permiso para verlo.

9.6 Registro de consultas

Como previamente indicamos, esta investigación busca mejorar el libro de visitas, ahora llamado registro de consultas, en el sentido en que, para la nueva versión del CHEM, se creen dos registros que proporcionen información más útil y precisa de la que muestra el libro de la primera versión, contabilizando únicamente cuando el corpus es consultado y no cuando un usuario ingresa al sistema. En este apartado precisaremos la forma en que ambos registros funcionan.

Registro de consultas de usuarios registrados: éste se integra por registros que se originan cada vez que un usuario realiza alguna búsqueda en el corpus. El enunciado anterior refleja la mejora que tiene, ya que ahora éste empieza a contabilizar a partir de que el usuario

realiza alguna consulta y no simplemente cuando el usuario ingresa. Por lo anterior, el libro de visitas deja de llamarse así para tomar el nombre de Registro de consultas.

El registro guardará la hora de la consulta del usuario, así como la referencia a los datos del mismo. Finalmente, el registro de consultas de usuarios registrados se muestra en la Figura 25**.**

Registro de consultas de usuarios registrados

Nombre	Correo	País	Institución	Consultas
Karla Perez	karlaperez@correo.com	mx	Facultad de Filosofia	15
Manuel Vega	manuel@correo.com	mx	II	15
Margarita Medina	margaritamedina@correo.com	mx	GIL	9
Samanta Lopez	samantalopez@correo.com	mx	Facultad de Filosofia y Letras	8
Gabriel Perez	gabrielperez@correo.com	mx	FES Zaragoza	5
Hugo Aguilar	hugoaguilar@correo.com	mx	Facultad de Ingenieria	2

Figura 25. Registro de consultas de usuarios registrados

Registro de consultas de usuarios anónimos. Este registro ha sido desarrollado para la nueva versión del CHEM con la finalidad de brindar mayor información del uso del corpus, ya que como hemos mencionado, el generador de concordancias puede ser utilizado, con ciertas limitantes, por usuarios no registrados.

El registro de consultas de usuarios anónimos se encarga de crear registros cada vez que algún usuario realiza alguna búsqueda. Tomando en cuenta que se desconocen los datos de los usuarios, en este libro se guarda la dirección IP, el país y un contador de accesos para cuando esa IP y país han sido registrados con anterioridad. La siguiente figura ilustra el registro de consultas de usuarios anónimos.

Registro de consultas de usuarios anónimos

País	IP	Consultas
Canada	134.190.57.83	1
France	157.23.124.51	2
Germany	188.1.14.65	1
Mexico	189.145.22.8	7
Mexico	189.145.235.120	2
Mexico	187.159.180.163	1
Mexico-UNAM	132.248.156.253	7
Utited States	184.128.67.45	4

Figura 26. Registro de consultas de usuarios anónimos

Cabe destacar que ambos registros únicamente pueden ser visualizados por el usuario de tipo administrador, debido a los datos personales de los usuarios que éste contiene, además de que la información que se muestra es útil únicamente para él.

9.7 AJAX

Anteriormente habíamos mencionado que AJAX es un conjunto de tecnologías que ayudan a responder a peticiones del servidor Web de manera asíncrona. Dichas peticiones pueden referirse a recepción o envío de información que el servidor procesa y manipula dependiendo de las necesidades del cliente, con la variante de que sólo se cambiará el aspecto de la parte de la página que lo requiere.

Ahora explicaremos cómo es que fueron integradas dichas tecnologías a la nueva versión del CHEM no sin antes mencionar las siguientes consideraciones para la integración de AJAX a la aplicación[36]:

- Para la mayoría de los enlaces utilizamos el controlador de eventos *onclick* para realizar llamadas asíncronas, debido a que a diferencia del comando *href* no muestra la dirección URL en la barra de estado del navegador[37].

[36] Queremos aclarar que los atributos o controladores de eventos de las etiquetas HTML que se utilizan para aplicar AJAX depende totalmente de las necesidades de la aplicación que se desarrolla.

- Preferimos utilizar el tipo *button* de la etiqueta *input*, ya que al utilizar AJAX no es necesario crear tipos *submit* que envíen las acciones al atributo *action* de, por ejemplo, un formulario, debido a que AJAX es quien se encarga de procesar los datos.

Para poder utilizar AJAX construimos un archivo de tipo Java Script, el cual contiene dos funciones principales, entre otras, para permitir actuar a las llamadas asíncronas, las funciones las exponemos a continuación:

- llamarAsincrono: esta función se utiliza únicamente cuando la parte que va a cambiar de la página se refiere a presentación, es decir, cuando se requiere del método GET.

- llamarAsincronoP: por su parte, esta función la utilizamos cuando es necesario enviar datos de una página a otra, cuando se requiere el método POST.

Posteriormente, comenzamos a crear los enlaces de la aplicación. En caso de que se requiera la función llamarAsincrono se construye de la siguiente forma:

<input type="button"
onclick="javascript:llamarAsincrono('contenido/inicio.jsp','sobrecontenedor');" class="cursor"
value="Cancelar" name="Cancelar"/>

Donde el controlar de eventos *onclick* indica que se está utilizando una función Java Script: *llamarAsincrono*, a la cual, en primer lugar, se le envía la referencia de la página que responderá al evento y, en segundo lugar, el *div* al que se enviará dicha respuesta, es decir, la parte de la página que cambiará su contenido.

En caso de que se requiera de utilizar la función llamarAsincronoP, la línea se construye como sigue:

<input type="button"
onclick="javascript:llamarAsincronoP('contenido/ValidarAccesoSistema.jsp','global',
this.parentNode)" class="cursor" value="Iniciar" name="Isesion"/>

[37] Esto no quiere decir que AJAX funcione únicamente con la propiedad onClick, también lo hace con onBlur, onFocus, etc., pero depende totalmente de las necesidades de la aplicación.

Donde, al igual que en la función anterior, *onclick* nos indica que se está utilizando una función Java Script *llamarAsincronoP ,*la cual recibe como primer parámetro la referencia de la página que responderá al evento, enseguida el *div* que recibirá dicha respuesta y por último se indica con this.parentNode que se enviarán todos los objetos HTML que se encuentran en la página.

Conclusiones

Este último capítulo está dedicado a exponer las conclusiones surgidas de la realización de esta investigación. Para ello, primero se hace una breve reseña de cada uno de los capítulos de la tesis. Luego son presentados los asuntos que quedaron pendientes de resolver; los nuevos problemas que surgieron después de concluir esta tesis, los cuales pueden ser base de otras investigaciones; y las posibles desventajas de las tecnologías y métodos utilizados. Posteriormente, una a una son revisadas las hipótesis planteadas al principio y reformuladas en caso de ser necesario, esto con el propósito de servir a investigaciones futuras. Después se exponen las ventajas y los logros obtenidos de acuerdo a los objetivos previamente formulados. Finalmente se puntualizan las conclusiones generales de esta investigación.

El primer capítulo se dedica a presentar la investigación realizada para entender conceptos como lingüística e ingeniería lingüística, con ellos también se exponen los diferentes niveles de análisis lingüístico: fonética/fonología, morfología, sintaxis, semántica, pragmática y discurso. Una vez hecho esto, se define el procesamiento de lenguaje natural y lingüística computacional para finalmente exponer sus áreas más representativas: recuperación de información, extracción de información y traducción automática.

En el segundo capítulo se estudia la lingüística de corpus. En él se presenta la definición de términos como la propia lingüística de corpus, corpus y corpus lingüístico. Después se exponen los diferentes tipos de corpus para posteriormente enfocarse en lo que es un corpus lingüístico electrónico. También son presentados algunos de los corpus lingüísticos electrónicos más importantes en nuestros días: el CREA y el CORDE de la Real Academia Española, el corpus del español de Mark Davis, el CLI y el CSMX del Instituto de Ingeniería de la UNAM, entre otros. Luego se detallan algunas herramientas de análisis de corpus lingüísticos como: listas de palabras y concordancias. Finalmente se incluyen algunos usos o aplicaciones lingüísticas y computacionales de un corpus.

El tercer capítulo se orienta a profundizar en el CHEM, es decir, se evidencian sus antecedentes y objetivos así como su arquitectura. Dentro de ésta se pone especial énfasis en el generador de n-gramas, en el generador de concordancias y en el proceso de incorporación de documentos. De igual forma se hace referencia a la base bibliográfica XML que sirve como contenedor para el total de documentos del corpus.

El siguiente capítulo se enfoca a la administración del CHEM. En primer lugar se hacen ver los dos principales problemas del registro de usuarios de la primera versión, proponiendo con ello una solución. Al respecto, se presenta el resultado del análisis e identificación de los tipos de usuarios para la nueva versión, y junto con estos se exponen los permisos que cada uno tendrá asociados. También, se propone el uso de una API para servir al registro de usuario y a la recuperación de sus contraseñas. En segundo lugar, se indican algunos aspectos relacionados con la administración de herramientas. Finalmente, en tercer lugar, se trata el tema de la administración de los documentos, que incluye la identificación de permisos y el desarrollo de una nueva herramienta: el visor de documentos. Para el visor, se estudian algunas herramientas para presentar los documentos y se hace una elección preliminar de una de ellas.

El capitulo quinto se orienta a uno de los temas medulares de este trabajo; en él se trata la investigación realizada para la creación de una nueva herramienta de análisis de corpus: un generador de estadísticas. Primeramente se define y se detallada la construcción de una tabla de contingencia, a partir de la cual se pueden calcular las medidas estadísticas. Luego son definidas las cuatro medidas estadísticas contempladas: información mutua (I), razón de semejanza ($-2\log\lambda$), prueba de independencia (X^2) y prueba de coligación de Yule (Y); también se incluyen algunas sus ventajas y desventajas. Al final se detallan algunas de las aplicaciones en las que estas estadísticas resultan de utilidad.

En el sexto capítulo, la seguridad en el CHEM, se describe, por un lado, la forma en que se espera que sea asegurada la información que proporcionan los usuarios al llevar a cabo su registro, y por otro lado, se plantean diferentes herramientas que pudieran ser útiles para proteger a los documentos que serán mostrados por el visor de documentos. Una vez expuestas las ventajas y desventajas que ofrece cada herramienta, se elige la que brinda mejores ofertas de seguridad y de conversión de documentos: Adobe Professional 9.

En el séptimo capítulo se propone la inclusión de una serie de tecnologías denominadas AJAX a la aplicación de administración y consulta del CHEM. Se define el concepto de dichas tecnologías y se utilizan algunos ejemplos para conocer un poco su funcionamiento.

En el siguiente capítulo, el octavo, se revisan los problemas que trae consigo el libro de visitas de la primera versión, se plantea la importancia de resolver dichos problemas y se propone la mejora del libro, primeramente justificando el cambio del nombre del libro y luego distinguiendo entre dos registros de usuarios.

Finalmente, el capítulo nueve se dirige a mostrar cómo es que cada nuevo aspecto fue integrado para crear la nueva versión del CHEM. En primer lugar, se describe cada una de las diferentes formas de administración que contendrá la nueva versión: administración de usuarios, administración de los documentos y administración de herramientas. Posteriormente se manifiesta la forma en que fue construido el generador de estadísticas, para que de él resultaran las medidas esperadas. Luego se expone la forma en que funciona el visor de documentos y se ilustra su resultado. Después, se trata la integración de los registro de consultas (previamente libro de visitas). También, se indica el resultado de utilizar Java Mail para envío de correo electrónico y finalmente se describe como fue integrado AJAX a la aplicación.

En cuanto al tema de los asuntos que quedaron pendientes por resolver puntualizamos algunos que bien podrían ser base de investigaciones posteriores:

- Desarrollar colocaciones, agrupaciones o listas de palabras. En este trabajo se desarrolló un generador de estadísticas que en combinación con el generador de concordancias (herramienta construida e integrada desde la primera versión del CHEM), podría resultar muy útil para construir con relativa facilidad nuevas herramientas de explotación del corpus, por ejemplo, colocaciones, agrupaciones o listas de palabras. El hacer nuevas herramientas resultaría de gran utilidad ya que incrementaría las posibilidades de obtener información del corpus.

- Construir rutinas para optimizar los registros de la base de datos del CHEM. Para el registro de usuarios, como se dijo, es enviado un correo que permite al usuario activar su cuenta; sin embargo, si el usuario hace caso omiso del correo, la base de datos del CHEM guarda el registro del usuario sin la cuenta activada. En este sentido, sería útil agregar rutinas que se encarguen de borrar periódicamente de la base de datos a

todos aquellos registros que no tienen el atributo *estado_usuario* como verdadero después de por ejemplo seis meses.

- Investigar otras opciones para identificación de IP. Si bien en este trabajo se investigaron algunos métodos para obtener la dirección IP de los usuarios que visitan o consultan el CHEM, el método que finalmente se eligió fue el desarrollado por un tercero debido a que resultó ser ideal para las necesidades de la aplicación; sin embargo, con este también se crea una dependencia del CHEM hacia otro sistema. Dicha dependencia pudiera no ser un problema siempre y cuando la aplicación que obtiene la IP siga funcionando, por ello resultaría importante desarrollar en un futuro un método que capture la IP, del cual se tenga cierta certeza de que no dejará de funcionar.

- HTTPS para transporte de información. Aunque esta tesis propone la inclusión de métodos para, de alguna forma, asegurar que la información que el usuario proporciona es correcta y no representa un peligro a la integridad de los datos, sería importante añadir un protocolo de seguridad que permita el transporte de los datos por un canal seguro, es decir, HTTPS.

- Métodos de protección para documentos. Se propone hacer una investigación, mucho más minuciosa de la que se hizo aquí, para encontrar métodos que ayuden a proteger los documentos. El hacerlo, bien podría aumentar la seguridad que se tiene de que los documentos no pueden ser plagiados.

Para poder cumplir con los requerimientos de la nueva versión del CHEM, que fueron base para este trabajo, se tuvieron que agregar nuevas tecnologías y métodos; sin embargo, cada una de ellas trae consigo desventajas. Para continuar, primeramente se tratan las desventajas de las tecnologías y posteriormente las de los métodos utilizados.

Java Script. La principal desventaja y quizá las más importante para la nueva versión del CHEM es el uso de Java Script debido a que puede ser activada o desactivada desde el navegador del usuario, y si esto sucede, la aplicación simplemente no podría funcionar, porque sería imposible enviar o recibir datos del servidor.

AJAX. Dentro de las desventajas de AJAX se encuentra su dependencia con Java Script, sin éste no funciona, así que si el usuario desactiva Java Script de su navegador también se estará negando la posibilidad de visualizar la página del CHEM. Otro desventaja importante es que AJAX no permite que el usuario navegue hacia atrás en las páginas que va consultando porque AJAX no genera un historial. También si la página se recarga, se muestra desde el principio y puede que no sea la página que se estaba consultando.

En cuanto a los métodos utilizados, las desventajas que se encuentran son las siguientes:

1. La más importante es que el método que se encarga de capturar de la IP de los usuarios fue desarrollado por un tercero, así que si la página no funciona temporal o indefinidamente, la aplicación de consulta del CHEM no puede ser visualizada. Lo anterior porque lo primero que hace la página del CHEM al ser cargada es revisar la IP del usuario para poder mostrarle las opciones correspondientes.

2. Otra desventaja es que, aunque se estudiaron varias herramientas para proteger a los documentos que se muestran en la nueva versión del CHEM y se identificó a la que ofreció mayores ventajas, se sabe que existen herramientas que se encargan de violar las restricciones a los documentos en casi cualquier formato, aunque éstas no son muy populares.

Para continuar se procede a revisar cada una de las hipótesis formuladas al principio de esta investigación.

Para la nueva versión del CHEM se construyó exitosamente un generador de estadísticas que calcula, haciendo uso del corpus, medidas como: información mutua, razón de semejanza, prueba de independencia y coeficiente de coligación de Yule. Estas medidas logran reflejar la relación que existe entre las palabras que surgen como resultado de las concordancias, además de que cada una tiene ventajas o desventajas sobre la otra dependiendo por ejemplo del número de datos que se tengan.

Con lo anterior se puede aseverar que la primera hipótesis resultó ser correcta. Pero correcta desde el punto de vista del CHEM, ¿qué quiere decir esto? Quiere decir que

efectivamente se puede afirmar que se agrega valor a las investigaciones que se puedan realizar, porque ahora el CHEM propone un mayor contexto para la investigación al no sólo mostrar resultados de concordancias (se recuerda que en la primera versión únicamente se contaba con un generador de concordancias como herramienta de análisis del corpus), sino distinguiendo medidas estadísticas precisas sobre el corpus (información mutua, razón de semejanza, prueba de independencia y coeficiente de coligación de Yule). Se quiere aclarar que al decir "valor a las investigaciones" no significa que la investigación que sobre el corpus se realicen tenga mejores resultados, eso depende totalmente del investigador. Lo que hace ahora el CHEM es brindar mayor información, lo cual hace el valor adicional.

Sin embargo, cabe la posibilidad de reformular la hipótesis, para no caer en confusiones, de la siguiente forma:

- El uso de las medidas estadísticas: información mutua, razón de semejanza, prueba de independencia y coeficiente de coligación de Yule permitirá que el CHEM muestre información sobre la relación entre palabras dentro de las concordancias.

La primera versión del CHEM no contaba con algún procedimiento elaborado para gestionar a los usuarios, éstos únicamente debían registrarse a través de un formulario y en seguida podían hacer uso del generador de concordancias. Derivado de lo anterior, en la nueva versión del CHEM se contempló la tipificación de usuarios con el objetivo de administrar los recursos del corpus y conocer el alcance del mismo. Ahora, aquellas personas que deseen hacer uso del corpus se registrarán y darán de alta su cuenta por medio de un correo electrónico para luego poder utilizar el corpus de acuerdo a los permisos que tenga asignado su tipo de usuario.

Una de las ventajas de tipificar a los usuarios es la posibilidad de concederles limitada o ilimitadamente el uso de las herramientas y materiales del corpus, según su tipo y de acuerdo a la conveniencia del CHEM. Esta idea, como se puede ver, está muy alejada de lo que era la administración en la primera versión, en la que además un usuario podía registrarse sin proporcionar todos sus datos. Ahora con la administración de usuarios del CHEM, el administrador puede gestionar a cada uno de los tipos de usuarios y permitirles o negarles el uso de las herramientas con relativa facilidad a través de diversas interfaces. Además, cada

vez que los usuarios realizan alguna consulta, inmediatamente ésta se registra en un registro de consultas que después puede servir al administrador para tomar decisiones.

Cabe resaltar que la administración de usuarios desarrollada en este trabajo, también contempla a aquellos usuarios que no se registran en el CHEM o que se encuentran dentro de Ciudad Universitaria pero que finalmente también pueden utilizarlo. De ellos, la aplicación de consulta obtiene su dirección IP para registrarla en el registro de consultas y proporcionar al administrador más información acerca del uso y alcance del CHEM. A los usuarios no registrados también se les pueden negar o conceder permisos porque forman parte del grupo de tipos de usuarios contemplados en la aplicación.

Con lo anterior, la segunda hipótesis de esta investigación se logra verificar. La tipificación de usuarios y los métodos construidos en este trabajo de investigación, logran ayudar al administrador a gestionar de mejor manera los permisos de las herramientas a las que puede tener acceso el usuario.

Para el registro de usuarios, esta tesis propuso el envío de un correo electrónico para que el usuario confirme y active su cuenta y con esto concluya su registro. La finalidad de enviar un correo electrónico es en primer lugar probar que el registro fue realizado por una persona y en segundo lugar que la dirección de correo electrónico existe.

La seguridad que se puede tener en los enunciados anteriores se encuentra en que es menos probable que exista un programa que además de realizar registros falsos también entre a una cuenta de correo, abra un correo y dé clic en un enlace. La mayor seguridad de que el registro puede ser correcto es porque si la dirección de correo no fuera correcta, nunca será enviado el correo electrónico y mucho menos el usuario podría acceder al CHEM. No por algo muchos sistemas han optado por el envío de correos electrónicos para confirmar registros. Por las líneas anteriores, la tercera hipótesis de esta tesis puede confirmarse debido a que no se aventura a decir que el método utilizado para confirmar y activar registros brindará total seguridad sino que agrega seguridad y confianza en buena media.

Dentro de cada uno de los diferentes formularios a los que el usuario, ya sea del tipo administrador o cualquier otro, proporciona información, este trabajo integra métodos que

permiten la validación de los datos, dicha validación se lleva a cabo tanto a nivel cliente, utilizando funciones Java Script como a nivel servidor a través de una clase en lenguaje Java.

Las funciones Java Script, se encargan tanto de verificar como de validar el formato de los textos que proporciona el usuario además de comprobar que todos los campos contengan datos. Por otra parte, los métodos construidos en Java se encargan de verificar que los datos proporcionados no son datos que pudieran afectar a la aplicación a nivel de la base de datos, es decir que estos no sean lo que comúnmente se conoce como inyecciones SQL.

Añadiendo métodos y funciones como las anteriores se prevé que la posibilidad de que el sistema acepte formularios con datos incompletos o erróneos será mínima, primero porque como se ha mencionado, la aplicación de consulta no puede funcionar si el usuario tiene desactivada la función de Java Script de su navegador, y segundo porque al tener Java Script desactivado le resultaría al usuario prácticamente imposible enviar datos a la aplicación; sin embargo, si lograra hacerlo, Java se encargaría de comprobar que ese esos datos no pueden afectar a la integridad del CHEM y a sus usuarios. Entonces, la cuarta hipótesis de esta investigación se confirma porque efectivamente Java Script y algunos métodos construidos en Java permiten generar certeza en que los datos proporcionados son completos y correctos en cuanto al formato y se puede pensar que son veraces porque cumplen con las características anteriores.

En este trabajo, como en su momento se señaló, se logró encontrar un método que permite la identificación de la dirección IP de los usuarios cuando éstos consultan el corpus. Es importante indicar que con ese método se logró discriminar el rango de aquellas direcciones IP que se encuentran dentro del espacio de Ciudad Universitaria de aquellas que se encuentran fuera. La identificación anterior permitió asignar ciertos recursos a la comunidad universitaria.

El método encontrado, aunque eficaz, porque cumple el objetivo de capturar la IP relativamente rápido, no brinda al CHEM la seguridad de que siempre estará cumpliendo su cometido porque está hecho por un tercero que se desconoce y que en cualquier momento podría decidir detener el funcionamiento de su aplicación. Por lo anterior, la quinta hipótesis podría bien reformularse de la siguiente forma:

Existe algún método eficaz, eficiente y confiable para identificar la dirección electrónica (IP) de los usuarios que utilicen el CHEM dentro o fuera de la UNAM.

De este modo se intentaría redefinir al método que se busque utilizar para la obtención de la dirección IP, esperando que en primer lugar no sea desarrollado por un tercero que genere incertidumbre y que además sea eficiente en cuanto a que cumplirá su objetivo sin utilizar más recursos.

También se logró integrar al CHEM un visor de documentos, que permite que el usuario pueda ver por completo los documentos que conforman el corpus y de los cuales son extraídas las concordancias. El usuario, dependiendo su tipo, tiene acceso a un número de textos, así por ejemplo, un usuario básico puede ver únicamente los documentos de los que el GIL no requiere tener permiso de publicación y un usuario administrador puede tener acceso a todos los materiales del CHEM.

Se resalta que, aunado a los resultados que provee el generador de concordancias y ahora también el generador de estadísticas, el investigador tiene a su disposición, con la nueva versión del CHEM, el texto completo del cual se generan las concordancias, con esto se le ofrece mayor información al mostrarle el contexto origen de su búsqueda. Con lo anterior, la sexta hipótesis puede confirmarse.

Esta tesis también contempló el mejorar el libro de visitas que había sido construido para la primera versión del CHEM, este trabajo consiguió mejorarlo en cuanto a que ahora la información que refleja tiene que ver con el uso que se hace del corpus y no con el número de visitas que este pueda tener. Además, al administrador del CHEM le resulta más útil saber cuándo es que el corpus es utilizado y no sólo el número de veces que es visitado.

Se indica que el libro de visitas en este trabajo cambió su nombre, para llamarse Registro de consultas, este registro se subdividió a su vez en dos grupos: registro de consultas de usuarios registrados y registro de consultas de usuarios anónimos. La primera mejora que se hace notar es que ahora se distingue también a los usuarios anónimos, es decir, aquellos que no se registran pero que si utilizan, limitadamente, las herramientas del corpus.

El registro de consultas de usuarios anónimos muestra al administrador la dirección IP de los usuarios, el país de origen de esa dirección IP y el número de consultas que realizan. Con esto el administrador puede saber, primordialmente, desde qué lugares y con qué frecuencia está siendo visitado el CHEM.

El registro de consultas de usuarios registrados proporciona al administrador datos de los usuarios cuando estos realizan alguna consulta en el corpus. La información que se puede ver en el registro es el nombre, el correo, el país y la institución a la que pertenece el usuario y el número de consultas que genera. Dicha información permite al administrador saber, al igual que el registro anterior, en qué lugares se está utilizando el CHEM y con qué frecuencia, además puede ayudar a que el administrador tome decisiones.

Los párrafos anteriores muestran que el mejorar el libro de visitas, ahora llamado registro de consultas, puede proporcionar información mucho más útil de la que en la primera versión se tenía. Por ello, la séptima hipótesis no se puede falsear.

Para confirmar la última hipótesis formulada en esta investigación se integraron una serie de tecnologías denominadas AJAX a la nueva aplicación de consulta del CHEM. Dichas tecnologías permiten la comunicación asíncrona con el servidor, es decir, se pueden realizar peticiones independientes del total del contenido, afectando únicamente la parte que lo necesita, de este modo el tiempo para que la respuesta en la interfaz esté disponible es relativamente bajo a comparación de si se recargara toda la página cada vez que se haga una solicitud.

Estas tecnologías se pueden notar principalmente cuando se utiliza el visor de documentos o el generador de concordancias debido a que el tiempo de espera es solo el tiempo que se lleva cargar el documento o generar las concordancias en lugar de además volver a recargar toda la página. Por lo anterior, la aplicación de AJAX, sí proporciona al CHEM una mayor capacidad de respuesta en cuanto a tiempo y presentación

Como se hizo notar, gran parte de las hipótesis fueron demostradas al haber desarrollado o encontrado las herramientas y métodos informáticos necesarios para cumplir con los objetivos. Por lo tanto y para concluir con la revisión de las hipótesis, se indica que se puede confirmar la hipótesis principal, punto de partida para el desarrollo de esta tesis.

Esto es, se desarrolló una herramienta de análisis (generador de concordancias), un visor de documentos, se construyó una administración para los usuarios y se dotó a la nueva versión de opciones para gestionarlos, además de la mejora de libro de visitas (registro de consultas) y la integración de tecnologías AJAX. Por lo anterior se remarca que desarrollando aplicaciones informáticas se facilitó la administración del CHEM así como se ayudó al análisis estadístico en su nueva versión. Entonces, de lo anterior se infiere que el desarrollo de aplicaciones informáticas ayuda a la administración y análisis de corpus lingüísticos electrónicos.

Para continuar con la presentación de estas conclusiones, se hablará sobre los objetivos planteados al comienzo de esta tesis. Es importante señalar que aunque para este trabajo se plantearon tanto objetivos prácticos como de investigación, únicamente se mencionará si se cumplieron o no los prácticos porque al hacerlo también se aludirá a los objetivos de investigación, los cuales fueron la base para el logro de los primeros.

Dada la investigación de las medidas estadísticas, se cumplió el primer objetivo al conseguir desarrollar una herramienta de análisis para la nueva versión del CHEM, la cual genera estadísticas a partir de las concordancias, que dejan conocer la relación que existe entre las palabras.

También se logró agregar un método para envió de correo electrónico que no sólo ayuda a que el usuario concluya su registro, sino que también resulta ser útil cuando el usuario necesita recuperar su contraseña.

Por otra parte, se cumplió con el objetivo de tipificar a los usuarios que harán uso del CHEM, así como el desarrollo de opciones administrativas para la asignación de herramientas y de permisos de documentos a cada uno de los tipos de usuarios. Además, también fue posible construir una interfaz para la creación de nuevos tipos de usuarios, considerando que estos también necesitarán ser gestionados.

Asimismo se consiguió, como se dijo en los objetivos, reconocer a aquellos usuarios que utilizan la aplicación de consulta del CHEM dentro de la UNAM, por medio de la identificación de su dirección IP. Al ser identificados los usuarios, fue posible tipificarlos y

posteriormente asignarles el uso de algunas herramientas con menos restricciones de las que tendrían estando fuera de Ciudad Universitaria.

Cumpliendo con el objetivo de validar la información que es proporcionada por los usuarios, se logró integrar métodos construidos en Java Script dedicados a validar que la información ingresada es correcta, completa y de cierto modo veraz.

El objetivo de mostrar los documentos completos a determinados tipos de usuarios se cumplió al encontrar una herramienta que permite cargar los textos en la aplicación de consulta, además fue posible asignar la posibilidad de verlos a determinados usuarios porque, como antes se mencionó, el CHEM integra opciones para gestionar sus herramientas. Con la herramienta que permitió mostrar los documentos también se consiguió aplicar restricciones de seguridad a los documentos del corpus, cumpliendo así con otro objetivo.

En cuanto al objetivo de optimizar el libro de visitas, con el fin de que se proporcione información útil y confiable acerca del alcance y uso del CHEM, podemos decir que se cumple dado que ahora se crean registros cada vez que el corpus es consultado y no visitado, como en su primera versión. Además tanto las validaciones de la información proporcionada en el registro como el método de identificación de IP ayudan a incrementar la confianza que se puede tener en el registro. Este objetivo también se cumple porque ahora se consideran las consultas que generan los usuarios anónimos.

Para terminar con la revisión de los objetivos prácticos, se indica que la aplicación de AJAX al CHEM logró proporcionar mayor capacidad de respuesta al no requerir que la página se cargue por completo cada vez que se hace una solicitud. En cuanto a la presentación y para las necesidades de la aplicación, AJAX resulta ser de gran utilidad porque los contenidos están en función de los tipos de usuario, así que sería difícil trabajar con frames, por ejemplo, porque su estructura es más rigurosa que la de los divs.

En lo que se refiere al objetivo principal de esta tesis se señala que tras la verificación del cumplimiento de los objetivos específicos podremos concluir que el objetivo se ha alcanzado satisfactoriamente debido a que en la nueva versión del CHEM se han corregido los errores de la primera versión. Asimismo se crearon nuevas herramientas: el visor de documentos y el generador de estadísticas. Con esto, se hace una aportación tecnológica al

CHEM con el fin de ayudar a que éste cumpla el objetivo principal para el cual fue creado, señalado en los primeros capítulos de este trabajo: "contar con herramientas de utilidad para investigadores y académicos, entre otros, que además pueda ser utilizado mediante una aplicación Web diseñada con herramientas para hacer exploraciones sobre el mismo".

A continuación se puntualizan los logros obtenidos junto con las ventajas de la realización de este trabajo:

- Se logró desarrollar una nueva herramienta de análisis del corpus llamada generador de estadísticas, éste calcula las medidas de: información mutua (I), razón de semejanza ($-2\log\lambda$), prueba de independencia (X^2) y prueba de coligación de Yule (Y), para cada una de las palabras que forman parte de las concordancias. La principal ventaja de este generador es que se convierte en una herramienta cuantitativa de análisis del corpus, misma que brindará información sobre las relaciones entre palabras. Otra ventaja es que facilitará el desarrollo de nuevas herramientas de análisis que pudieran ser integradas en un futuro.

- Se integró un método que permite el envío de correo electrónico, útil para la confirmación del registro y para la recuperación de contraseñas. La ventaja que esto representa es que ahora la confirmación permite asegurar que es una persona quien realiza el registro. Además, esto ayuda a optimizar la base de datos, en cuanto a que no se activan cuentas cuando el correo electrónico es falso o no se concluye el registro.

- Se logró, además de tipificar a los usuarios, desarrollar herramientas que permiten asignarles permisos sobre los contenidos del CHEM (herramientas y documentos), es decir, se logra administrarlos. Sobre este punto existen varias ventajas en comparación con la primera versión. La primera de ellas es precisamente que ahora el CHEM cuenta con una administración mucho más exhaustiva al distinguirse tipos de usuarios. La segunda ventaja es que se permite la creación de nuevos tipos de usuarios, y la tercera, que a éstos es posible asignarles permisos sobre el uso de las herramientas del CHEM.

- Se consiguió reconocer a aquellos usuarios que hacen uso de la aplicación de consulta dentro de la UNAM, al obtener por medio de un método su dirección IP. Con ello también se logra asignarles a esos usuarios algunos beneficios para uso del corpus. La ventaja de esto reside en los beneficios que reciben los usuarios por estar dentro de la UNAM o pertenecer a la comunidad universitaria.

- También se logró construir métodos que se encargan de validar la información y se aseguran de que ésta no representa una amenaza a la seguridad de la base de datos. La principal ventaja que estos métodos conllevan, es que se disminuye el riesgo de afectar al sistema introduciendo datos erróneos, falsos o peligrosos.

- Ahora, también es posible mostrar los documentos completos a los usuarios por medio de un visor de documentos. Asimismo, se logró aplicar restricciones a cada uno de los documentos que se presentan para disminuir o evitar la posibilidad de que éstos puedan ser plagiados por personas malintencionadas y con el fin de mantener el respeto a los derechos que sobre ellos tiene el GIL. En este punto cabe resaltar que son muy pocos los corpus que muestran a los usuarios sus documentos, comúnmente solo muestran los resultados de las consultas que se hacen sobre ellos. Con lo anterior la ventaja es que se brinda más información a los investigadores.

- El registro de consultas, antes libro de vistas, logra mostrar información más confiable que en la primera versión, al hacer uso en el registro de usuarios de métodos que validan la información que se proporciona y del método encontrado para obtener la dirección IP de los usuarios. La ventaja de esto es el conjunto de oportunidades que puede tener el administrador del CHEM al tener una base confiable para conocer el alcance o uso de la aplicación, y con la cual puede tomar decisiones.

- Finalmente, se logró aplicar AJAX a la nueva versión del CHEM sin que esto representara un conflicto con la herramienta que ya se había desarrollado en la primera versión: el generador de concordancias; por el contrario, AJAX resulta ser una ventaja porque ahora con las nuevas herramientas se puede reducir el tiempo de respuesta al hacer llamadas asíncronas.

Para terminar con la exposición de las conclusiones de esta tesis, se presentan una serie de conclusiones generales.

Con el desarrollo de esta tesis se ha podido demostrar que la informática es una técnica que no sólo se puede aplicar a aéreas administrativas si no que también es posible aplicarla a un área del conocimientos como lo es la Ingeniería Lingüística, específicamente se logró adaptar al CHEM para ayudarlo a cumplir con sus necesidades de información.

Una de las aportaciones más destacada de esta tesis fue la creación de una nueva forma de administración para el CHEM, con la cual un usuario de tipo administrador tiene nuevas posibilidades como la de asignar o denegar permisos a los tipos de usuarios para el uso de herramientas o consulta de documentos, la de crear nuevos tipos de usuarios o permisos de los documentos, y la de integrar herramientas para su gestión; todo esto mediante el uso de interfaces que simplifican el proceso.

Otras aportaciones, no menos importantes que la anterior, es la creación de una nueva herramienta para la explotación del corpus: un generador de estadísticas, así como el desarrollo de un método que permitió crear un visor de documentos completos útil para proporcionar mayor cantidad de información a los usuarios. Éstas son una gran aportación ya que ellas podrán dar valor a las investigaciones hechas con el CHEM, debido a que ahora el investigador cuenta con más contexto para la realización de sus investigaciones.

Finalmente, cabe resaltar que las herramientas y métodos desarrollados a lo largo de esta investigación, han sido sólo aportaciones a un proyecto tan grande como puede ser el CHEM y sólo se han sentado algunas bases para el futuro desarrollo de nuevas herramientas o de una administración mucho más elaborada de la que aquí se propone.

Referencias Bibliográficas

BABIN Lee. Beginning Ajax with PHP. Apress. Estados unidos. 2007. 255 pp

BIBER Douglas; CONRAD Susan y REPPEN Randi. Corpus Linguistics Investigating Language Structure and Use. Cambridge University Press. Reino Unido. 2000, 2002. 300 pp.

Blecua, J. M. et al. (eds). Filología e informática. Nuevas Técnicas en los estudios filológicos. Barcelona: Editorial Mileno y Universidad Autónoma de Madrid. 1999. 52 pp.

Bolshakou y Gelbukh, 2004: 63-66

Chavarría Laura. "Procesamiento de Corpus Lingüísticos mediante el uso de bases de datos relacionales". Tesis de Licenciatura. Ciudad Universitaria. UNAM. 2008. 83 pp.

CHOPRA, Vivek... [et al.]. Profesional Apache Tomcat 5. Anaya multimedia. España. 2005. 605 pp

Gelbukh y Sidorov, 2006: 15-16

Kageura, Kyo. "Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences". *Journal of Quantitative Linguistics*, 1999. 41 pp.

Magaña López, Lisbeth et al. "Sistemas de extracción de información. Soluciones informáticas organizacionales basadas en datos no estructurados". Tesis de licenciatura. Ciudad Universitaria. UNAM. 2006. 255 pp.

MASON Oliver. Programming for Corpus Linguistics. How to Do Text Analysis with Java. Edinburg University Press. Edimburgo, Reino unido. 40-44 pp.

McEnery, Tony. "What is a corpus?" en R. Mitkov (ed.), The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford. 2003, pp. XXX – XXXX.

Medina Alfonso. "Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el *Corpus del Español Mexicano Contemporáneo*". Tesis de doctorado. El Colegio de México A. C. México. 2003. 464 pp.

Medina Alfonso; Méndez Carlos. “Arquitectura del Corpus Histórico del Español de México (CHEM)”, México, 2006, pp. 7.

Medina A., Sierra G., Garduño G. (2004). “Herramientas de análisis para el Corpus Lingüístico en Ingeniería” México, 2004. 8 pp.

MITKOW Ruslan. The Oxford Handbook of COMPUTACIONAL LINGUISTICS. Oxford University Press. Gran Bretaña. 2003. 784 pp.

Sinclair, 1994:4

Referencias WEB

(WEB, 01) II. Instituto de Ingeniería UNAM. Corpus Histórico del Español de México. http://www.iingen.unam.mx/txtlstvw.aspx?LstID=8ab5b49e-1265-4b3e-b925-7126b66fa36f>. [Consultada: 27 de febrero de 2009].

(WEB, 02) GIL. Corpus Lingüístico del Español. <http://www.iling.unam.mx/index.php?ID_HIST_INFORMACION=53>. [Consultada: 27 de febrero de 2009].

(WEB, 03) Centro Virtual Cervantes. La ingeniería lingüística en España. Ingeniería lingüística e industrias de la lengua. <http://cvc.cervantes.es/lengua/anuario/anuario_98/llisterri/llisterri_01.htm>

(WB, 04) Sidorov, Grigory. Problemas actuales de la lingüística computacional [en línea]. Revista Digital Universitaria (RDU). Vol.2, No.1 (Marzo 2001). http://www.revista.unam.mx/vol.2/num1/art1/. [Consultada: 17 de marzo de 2009].

(WB, 05) Sierra, Gerardo. Lingüística de corpus [en línea]. Introducción a la lingüística de corpus. <http://www.iling.unam.mx/CursoCorpus/1_1_Definicion.html>. [Consultada: 23 de marzo de 2009].

(WB, 06) Real Academia Española. http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=corpus. [Consultada: 25 de marzo de 2009].

(WB, 07) Sierra, Gerardo. Lingüística de corpus [en línea]. Tipología y Clasificación de corpus. <http://www.iling.unam.mx/CursoCorpus/1_2_Clasificacion.html>. [Consultada: 7 de abril de 2009].

(WB, 08) Sierra, Gerardo. Lingüística de corpus [en línea]. Clase1_Características. <http://www.iling.unam.mx/CursoCorpus/1_1_Definicion.html>. [Consultada: 15 de abril de 2009].

(WB, 09) Sierra, Gerardo. Lingüística de corpus [en línea]. Descripción de corpus existentes. <http://www.iling.unam.mx/CursoCorpus/1_3_Existentes.html>. [Consultada: 18 de abril de 2009].

(WB, 10) Mercedes Sánchez: CREA. Corpus de Referencia del Español Actual [en línea]. Departamento de banco de datos de la Real Academia Española, 2002. <http://www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos55A.pdf>. [Consultada: 21 de abril de 2009].

(WB, 11) REAL ACADEMIA ESPAÑOLA. CORDE disponible en: http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/B4E26FC2520104D8C125716400455C06?OpenDocument. [Consultada: 28 abril de 2009].

(WB, 12)DAVIES Mark. Un corpus anotado de 100.000.000 palabras del español histórico y moderno 2004 [en línea]. Procesamiento del Lenguaje Natural. No.29 (2002). <http://www.sepln.org/revistaSEPLN/revista/29/29-Pag21.pdf>. [Consultada: 28 abril de 2009].

(WB, 13) British National Corpus. What is the BNC? http://www.natcorp.ox.ac.uk/corpus/index.xml. [Consultada: 30 abril de 2009]

(WB, 14) INSTITUT UNIVERSITARI DE LINGUISTCA APLICADA. Proyecto Corpus Corpus textual especializado plurilingüe. http://www.iula.upf.edu/corpus/corpuses.htm. [Consultada: 30 abril de 2009]

(WB, 15) Corpus textuales en español. CUMBRE. http://homepage.mac.com/joaquim_llisterri/language_resources/lang_res/Corp_text_esp.html. [Consultada: 29 abril de 2009].

(WB, 16) Sierra Gerardo. Lingüística de Corpus. Concordancias. http://www.iling.unam.mx/CursoCorpus/4_2_Concordancias.html. [Consultada: 29 abril de 2009].

(WB, 17) Instituto de Ingeniería de la UNAM. Corpus Histórico del Español de México. http://www.iingen.unam.mx/txtlstvw.aspx?LstID=8ab5b49e-1265-4b3e-b925-7126b66fa36f. [Consultada: 2 julio de 2009].

(WB, 18) Grupo de Ingeniería Lingüística. Corpus Histórico del Español de México. http://www.iling.unam.mx/index.php?ID_HIST_INFORMACION=53. [Consultada: 8 agosto de 2009].

(WB, 19) Sun Microsystems. Sun Developer Network (SDN). http://java.sun.com/products/javamail/. [Consultada: 5 noviembre de 2009].

(WB, 20) iText. iText PDF: your Java-PDF library. <http://itextpdf.com/>. [Consultada: 18 noviembre de 2009].

(WB, 21) Adobe. Macromedia Flash Paper 2. <httphttp://www.adobe.com/la/products/flashpaper/productinfo/overview/>. [Consultada: 6 diciembre de 2009].

(WB, 22) Larson, Erik. Flash paper 2 Feature Tour. <http://www.adobe.com/products/flashpaper/productinfo/features/brz_tour/>. [Consultada: 4 diciembre de 2009].

(WB, 23) Adobe. Macromedia Flash Paper 2. <http://www.adobe.com/la/products/flashpaper/>. [Consultada: 6 diciembre de 2009].

(WB, 24) Adobe. Adobe Acrobat 9 Pro. Disponible en: <http://www.adobe.com/es/products/acrobatpro/features/>. [Consultada: 6 diciembre de 2009].

(WB, 25) Ejemplos java y C/Linux. Enviar un correo con JavaMail. <http://www.chuidiang.com/java/herramientas/javamail/enviar-correo-javamail.php>. [Consultada: 12 febrero de 2010].