Revista Digital Universitaria
1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

ARTÍCULO

EL CORPUS HISTÓRICO DEL ESPAÑOL EN MÉXICO

Alfonso Medina Urrea y Carlos Francisco Méndez Cruz

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

El Corpus Histórico del Español en México (CHEM)

1.1.Antecedentes

Los corpus lingüísticos constituyen uno de los tipos más prominentes de recursos digitales de uso en las humanidades. En la tradición lingüística los corpus se conocen como muestras textuales de diversas naturalezas, escritos u orales, representativos de alguna lengua, área temática, género literario, registro sociolingüístico, lenguaje de especialidad, etcétera. Hoy en día, los corpus lingüísticos son irremediablemente electrónicos. En México, la compilación de corpus electrónicos se inició en los años setenta, antes de la era de Internet. El primer corpus electrónico en español, el Corpus del Español Mexicano Contemporáneo de El Colegio de México, se constituyó como la base estadística de la nomenclatura del Diccionario del Español en México. Luego, con el advenimiento de Internet se han hecho disponibles al mundo los corpus de la Real Academia Española (CORDE y CREA), el Corpus del Español de Mark Davies y El Corpus Histórico del Español en México (CHEM). Este último, fue desarrollado en el Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería de la Universidad Nacional Autónoma de México, foco del presente artículo.

Los corpus son un recurso fundamental en las investigaciones lingüísticas, en el desarrollo de herramientas de procesamiento de lenguaje natural y en la construcción de otros recursos lingüísticos, como son los diccionarios, lexicones, etcétera. Por todo esto, en el GIL se han desarrollado herramientas computacionales de extracción automática de términos y definiciones (para la lexicografía y terminología computacionales), con el afán de apoyar de forma decisiva la labor de lexicógrafos y terminólogos. En este contexto, en el GIL se han abierto proyectos para elaborar corpus de diversas áreas, como la Ingeniería, los Contextos Definitorios (fragmentos textuales que incluyen en su interior un término y su definición) y las Sexualidades en México, lo que permitirá el estudio de documentos sobre las áreas de sexualidad y sexología y la futura creación automática de diccionarios electrónicos.

1.2.Definición

El CHEM es un corpus diacrónico, esto es, que comprende varios estados de la lengua (del siglo XVI al siglo XX) y que, durante los últimos seis años, nos ha permitido realizar investigación aplicada sobre la constitución de corpus lingüísticos electrónicos. Después de largo tiempo de investigación y desarrollo computacional, mediante el patrocinio de la DGAPA y el CONACyT, el CHEM está en línea en la dirección http://www.coprus.unam.mx/chem/. La

Figura 1 muestra la página inicial de su interfaz de consulta.

Figura 1. Página inicial de la interfaz de consulta del CHEM.

Este corpus incluye, por un lado, una colección de documentos producidos en la Nueva España y el México independiente, dispersos en distintos géneros textuales, y, por otro, las herramientas para explorar y analizar dicha colección. De hecho, busca apoyar las tareas de investigación de filólogos, lingüistas, historiadores, y todo aquel interesado en la cultura novohispana y del México de los últimos siglos. Por supuesto que el carácter diacrónico del CHEM lo hace un recurso especialmente útil en estudios de variación lingüística temporal, es decir, fenómenos que tienen que ver con los cambios que ha sufrido el español a través del tiempo.

1.3.Los documentos del CHEM

Como ya se dijo, el CHEM incluye documentos de diversos géneros textuales, pero, además, de distintas áreas temáticas, diferentes lugares, registros de lengua y distintos tipos de hablantes. Al incluir documentos de diversas zonas geográficas, podemos resaltar que puede ayudar a estudios dialectales, como son los cambios que sufre el español en distintos lugares. Respecto a la inclusión de documentos clasificados por el registro de lengua utilizado, es posible caracterizarlos según su "formalidad" o "falta de formalidad" de la escritura o habla de una persona o grupo de personas. Así, no es igual la expresividad lingüística encontrada en textos cuidados de carácter científico o académico, que en textos de uso común (registro estándar) o en textos familiares de esencia más íntima o personal. Todos ajustamos nuestra manera de hablar de acuerdo con los diversos ámbitos sociales en los que nos desenvolvemos. Si bien

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

estos no son todos los criterios en los que se podrían clasificar los documentos del CHEM, son con los que arrancó la versión actualmente en línea del mismo.

Con todo esto en mente, surge la pregunta: ¿es posible lograr equilibrio y representatividad en el CHEM? Este es uno de los aspectos de investigación al que el grupo de desarrollo del CHEM está abocado y en el cual ya tenemos avances. Una posible solución podría ser la que se utilizó para el Corpus del Español Mexicano Contemporáneo (CEMC) de El Colegio de México. Su estrategia fue la selección aleatoria y automática de fragmentos de los documentos hasta lograr que cada documento contara con la misma cantidad de ocurrencias de palabras. Esa estrategia permitió que ningún documento estuviera sobre representado, por tener mayor número de ocurrencias de palabras en comparación con otro. Sin embargo, proponemos el uso de frecuencias corregidas, basadas en un tamaño virtual del corpus. Esta idea se está trabajando y los resultados serán incluidos en la próxima actualización del CHEM.

1.4.El etiquetado de los documento

Preservar las características originales de los documentos fuente del CHEM. En la versión electrónica fue un requerimiento fundamental de este proyecto. De esta manera se buscó un formato adecuado para conservar las características ortográficas y de organización textual de los documentos, así como la información que fue agregada durante el trabajo de paleografía, hecho a partir del documento en papel, principalmente en los documentos más antiguos.

Pero conservar estas características no era el único requerimiento. También era necesario brindar una variedad de tipos de búsquedas, con el fin de hacer más útil el corpus. Se decidió que los tipos de búsquedas fueran, por: ortografía actual, ortografía original, lema, transcripción fonológica y categoría gramatical. Hasta el día de hoy, sólo las tres primeras están disponibles, pero para la próxima actualización del CHEM se incluirán más. Dejaremos la explicación de cómo realizar estas búsquedas para la sección dedicada a las herramientas del CHEM. Por ahora, sólo explicaremos lo necesario para entender cómo se les dio formato a los documentos del corpus.

Supongamos que queremos recuperar algunos ejemplos textuales en documentos del siglo XVI, en los que se use el verbo «decir». Según los documentos del CHEM, existen al menos las siguientes variantes ortográficas para este verbo: «desir», «dezir» y «decir». Además existe una variante que indica que esta palabra no estaba escrita de forma completa, por lo que el transcriptor, que hizo el trabajo de paleografía, reconstruyó la palabra de acuerdo al contexto en que aparecía y marcó con letra cursiva el segmento faltante: «decir». Nótese que el segmento ~*eci*~ está en letra cursiva y es por tanto una reconstrucción. Otro ejemplo sería el del verbo «vivir», para el que existen las siguientes variantes ortográficas: «vibir», «vivir», «bibir» y

«bj/6vir». Nótese cómo la última palabra no sólo cambia de ortografía, sino que también está separada por una marca de fin del línea (/), y en este caso además con marca del número de línea siguiente (/6).

Era imprescindible entonces encontrar un formato que permitiera codificar información agregada por el transcriptor a cada palabra y asociar la ortografía moderna con la antigua. Este formato nos lo brindó el lenguaje XML (*eXtensible Markup Language*). Este es un lenguaje para marcado (etiquetado) de documentos, que permite codificar dentro de los mismos información adicional. Además, brinda la posibilidad de estructurarlos de acuerdo a una convención elaborada por el investigador.

El lenguaje XML se basa en la marcación o etiquetado de elementos relevantes en los documentos. Como ya decíamos, cada marca o etiqueta es definida arbitrariamente por el investigador y se pone entre paréntesis angulares (<etiqueta>). Generalmente habrá una etiqueta de apertura y una de cierre, ésta última identificada porque antes del nombre lleva una diagonal (</etiqueta>).De esta manera cada palabra del CHEM fue etiquetada automáticamente con la etiqueta <g>. La palabra «desir» quedó codificada de la siguiente manera: <g>desir</g>. Para el caso de las reconstrucciones se utilizó la etiqueta . Así, la palabra reconstruida «decir» quedó codificada de la siguiente forma: <g>decir</g>.

En el caso de asociar distinta ortografía a las palabras, nos aprovechamos de que cada etiqueta puede tener atributos, esto es, un mecanismo que permite asociar distintos valores a una etiqueta. Para el caso de la ortografía modernizada o normalizada, se incluyó en cada etiqueta <g> un atributo n="" que permitió la asociación requerida. Entonces, la palabra «desir» quedó codificada de la siguiente manera: <g n="decir">desir</g>. Ahora recordemos el caso peculiar de las palabras cortada por marcas de renglón, como el caso expuesto arriba de «bj/6vir». Para salvar estos problemas, incluimos otro atributo a la etiqueta <g>, que tuviera la ortografía original, pero sin las marcas que separaban la palabra. Este atributo se llamó o="". De esta forma la palabra «bj/6vir» quedó codificada de la siguiente manera: <g n="vivir" o="bjvir">bj/6vir</g>. Nótese el uso del atributo n="" para la ortografía moderna y el uso del atributo o="" para la ortografía original.

Finalmente, las búsquedas por lema exigían que una palabra estuviese asociada a sus variantes morfológicas. Un lema es precisamente la forma de palabra que representa al conjunto de variantes. En el caso de los verbos es el infinitivo y en el caso de los sustantivos, el masculino singular. Entonces, cuando se realizara la búsqueda del lema «decir», el CHEM debía responder con las distintas formas verbales conjugadas, atestiguadas en el corpus: «dixese», «dixeron», «dizen», «Dezían», «dixesen», entre otras. Para lograr este tipo de búsqueda, cada variante morfológicas estaría asociada a su lema «decir» a través de un nuevo

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

atributo l="". Para ejemplificar cómo quedaron codificada estas palabras, ponemos sólo el etiquetado de la primera: <g n="dijese" l="decir">dixese</g>.

El lenguaje de etiquetado XML también nos ayudó a incluir información descriptiva de los documentos. Entre tal información se encuentra: título del documento; lugar y fecha de impresión; los criterios del documento, como: género literario, zona geográfica, área temática y datos del hablante; y además los datos de las personas que etiquetaron el documento. Esta información se incluyó en lo que llamamos el encabezado del documento. El cuerpo del documento, por su parte, está formado por el texto en sí, en donde se encuentran las palabras etiquetadas con la etiqueta <g>, tal y como se explicó arriba. A este cuerpo se le agregaron algunas otras etiquetas de carácter estructural, que nos permitieron marcar secciones y títulos de sección, entre otras.

1.5.Arquitectura de cómputo

Como puede notarse, el CHEM se forma de un conjunto de documentos XML que incluyen información lingüística y descriptiva de cada uno de ellos, mediante etiquetas definidas por el grupo de desarrollo del corpus. Este conjunto de documentos es procesado computacionalmente para facilitar y agilizar las consultas que el usuario hace desde la interfaz de consulta. Inspirados en la arquitectura computacional del Corpus del Español de Mark Davies (http://www.corpusdelespanol.org/), decidimos utilizar bases de datos relacionales para manejar este conjunto de documentos. La figura 4 muestra de forma esquemática esta idea.

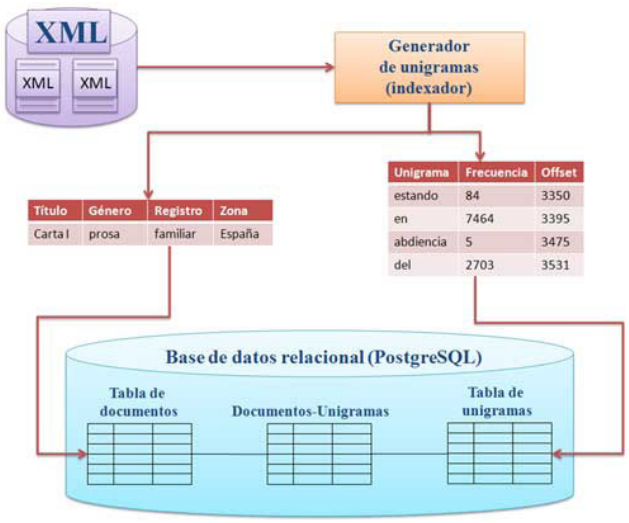


Figura 4. Uso de bases de datos relacionales para el manejo de los documentos del CHEM

Como puede apreciarse en la figura 4, cada documento XML del corpus es procesado por el

generador de unigramas (indexador). Este programa tiene el objetivo, por un lado, de extraer del encabezado de cada documento sus datos descriptivos, como: título, género literario, registro y zona geográfica. Estos datos son almacenados en la base de datos relacional en una tabla de documentos. Por otro lado, extrae cada unigrama o palabra de cada documento con su posición en bytes en el mismo (*offset*). Adicionalmente, acumula la frecuencia total del unigrama en todo el corpus.

El unigrama, frecuencia y *offset* son almacenados en una tabla de unigramas, que se vincula con la tabla de documentos para registrar la aparición de un unigrama en un documento determinado. Este proceso generador de unigramas se hace una sola vez para indexar todo el corpus y tener los registros listos para las búsquedas. Contar con la posición de cada palabra (unigrama) en el documento, nos permite agilizar las búsquedas en los mismos. La figura 5 muestra de manera esquemática el uso de la base de datos relacional en el proceso de búsquedas de palabras en el CHEM.

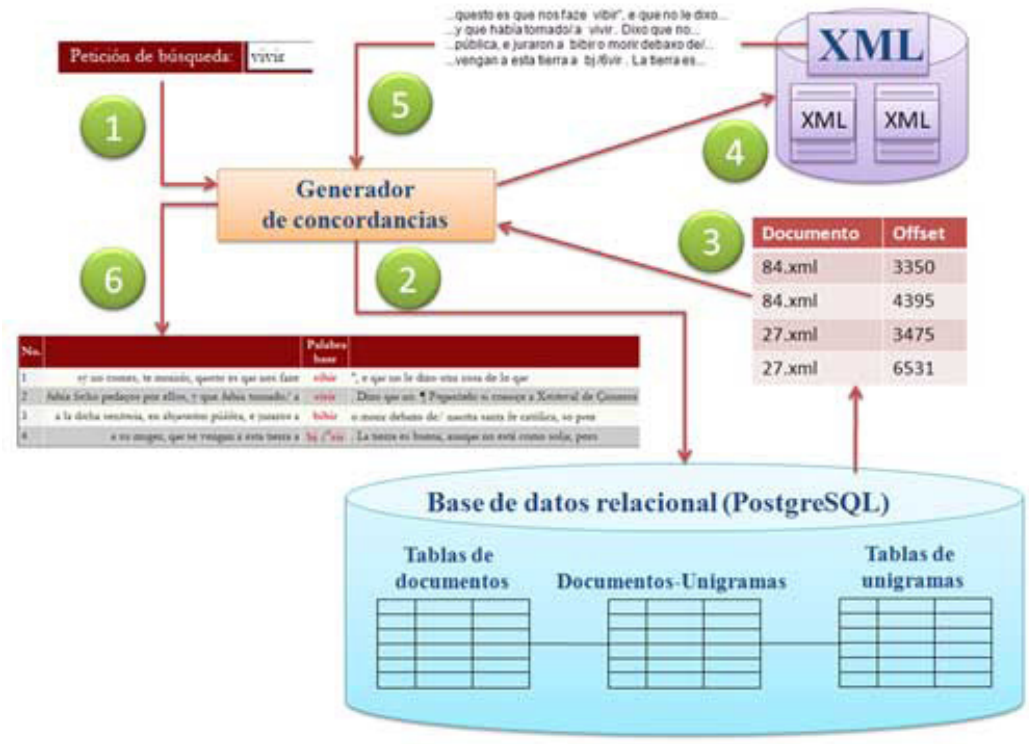


Figura 5. Uso de la base de datos relacional en el proceso de búsqueda de palabras en el CHEM.

Hemos incluido seis pasos en la figura 5, con el fin de facilitar la explicación del proceso de búsqueda de palabras en el CHEM. Todo el proceso inicia cuando el generador de concordancias recibe una petición de búsqueda por parte del usuario del corpus. (1). Una concordancia es un fragmento textual que incluye la palabra de búsqueda y cierto número de palabras a su izquierda y a su derecha. En el apartado de herramientas del CHEM hablaremos

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

un poco más sobre las concordancias. El generador de concordancias realiza una petición a la base de datos relacional para buscar la palabra. (2) La base de datos relacional regresa al generador de concordancias la lista de nombres de documentos XML, en donde se encuentra la palabra y sus *offsets*, es decir, las posiciones en las que ésta aparece. (3) Con esta información, el generador de concordancias localiza la palabra de manera directa en los documentos XML (4) y extrae las concordancias en las que aparece la palabra. (5) El generador de concordancias, entre otras cosas, da formato a estas concordancias y las muestras al usuario que hizo la petición. (6)

1.6.Los tipos de usuarios del CHEM

Con el fin de brindar beneficios adicionales a los usuarios más cercanos al corpus, como el uso no restringido de ciertas herramientas de análisis, decidimos establecer tres tipos de usuarios del CHEM. Por un lado están los usuarios registrados, es decir, aquellos que llenan un formulario de registro para proporcionarnos algunos datos. Estos usuarios tienen acceso casi ilimitado a las herramientas del corpus desde cualquier parte del mundo con sólo iniciar sesión mediante su correo electrónico y una contraseña que el mismo usuario determina. Llevar a cabo este registro también nos permite comunicarles avisos sobre la nueva información y funcionalidades incluidas en las actualizaciones del CHEM.

El segundo tipo de usuarios son los que acceden desde una red de computadoras que el CHEM detecta como privilegiada, como es la red UNAM. Esto nos permite dar ciertos privilegios a una comunidad académica sin necesidad de ser usuarios registrados, aunque tales privilegios serán menores que los asignados a estos últimos. El tercer tipo de usuarios del CHEM son lo que llamamos usuarios anónimos. Éstos son los que acceden al corpus y realizan búsquedas sin estar registrados ni ser parte de la red de computadoras que mencionábamos antes. Este tipo de usuarios tienen acceso a casi todas las funcionalidades del CHEM, pero siempre sus privilegios serán menores que los de los otros tipos de usuarios.

Creemos pertinente mencionar que el registro al CHEM es gratuito y no lleva demasiado tiempo, por lo que esperamos contar con más amigos del CHEM con los que podamos, en un futuro, interactuar para hacer crecer este corpus.

1.7.Registros de consultas al CHEM

El CHEM también cuenta con registros de las consultas que hacen los usuarios. Existen dos tipos, uno de usuarios anónimos y otro de usuarios registrados. El primero nos permite conocer de qué países nos visitan y cuántas consultas han hecho. Una muestra de este registro se puede ver en la figura 6. Como puede notarse, nos han visitado de diversos países de prácticamente todos los continentes.

Registro de consultas de usuarios anónimos

País	IP	Consultas
Anonimo	0.0.0.0	24
Argentina	201.216.244.197	9
Brazil	189.59.195.96	23
Brazil	187.114.22.160	14
Brazil	187.114.18.103	8
Brazil	200.139.85.70	3
Brazil	189.93.144.84	2
Czech Republic	195.113.21.61	2
Czech Republic	195.113.21.157	2
Denmark	95.209.218.146	1
France	86.71.43.197	5
Germany	92.50.108.31	3
Japan	122.17.56.82	2
Mexico	200.52.255.199	77
Mexico	189.146.201.158	31
Mexico	148.231.188.195	27
Mexico	189.146.170.112	24
Mexico	189.242.97.83	23
Mexico	201.137.97.182	22
Mexico	201.137.109.59	20
Mexico	132.247.245.162	19
Mexico	189.180.59.191	18

Figura 6. Muestra del registro de consultas de usuarios anónimos.

Como decíamos, el CHEM también registra las consultas de los usuarios registrados, pero no ponemos una muestra para no publicar los correos electrónicos de éstos. Sólo mencionaremos que contamos con usuarios registrados de diversas universidades tanto del país (Universidad Autónoma del Carmen, Universidad Autónoma del Estado de México, Benemérita Universidad Autónoma de Puebla, Universidad Autónoma Metropolitana plantel Azcapotzalco y Universidad Autónoma de Baja California) como del extranjero (Universidad Pompeu Fabra, Universidad de Utrecht, Universidad de las Palmas de Gran Canaria y Universidad de Cambridge).

2.Las herramientas del CHEM

2.1.El generador de concordancias

Una concordancia se forma de una palabra base y su contexto textual. Este contexto está determinado por una ventana de caracteres o palabras a la izquierda y derecha de la palabra

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

base, a veces llamada palabra pivote. También una concordancia se acompaña de la referencia al texto de donde se obtuvo el fragmento. Así, la figura 7 muestra las concordancias de la palabra «inquisición» obtenidas del CHEM.

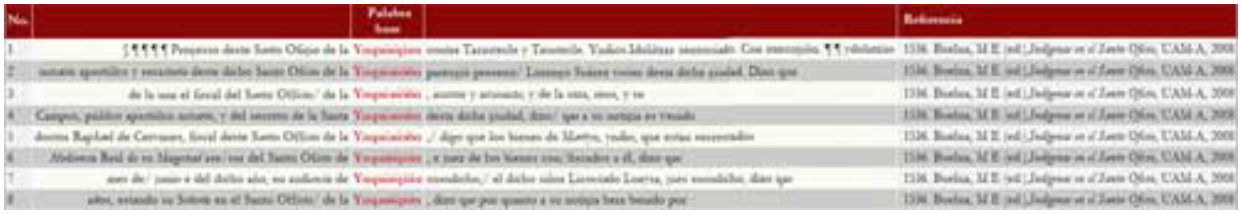


Figura 7. Concordancias de la palabra «inquisición» obtenidas del CHEM.

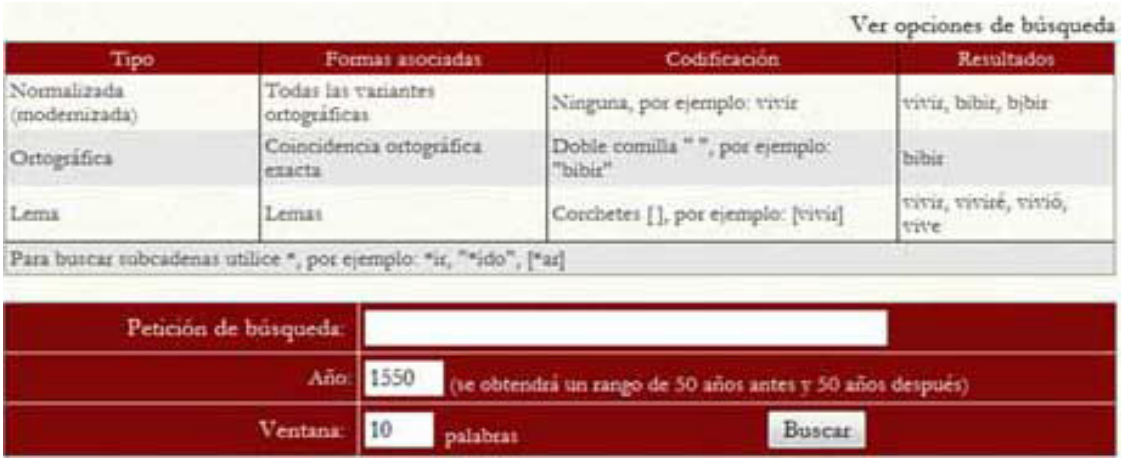
La gran mayoría de los corpus electrónicos disponibles en la red incluyen una herramienta para generar concordancias. Cada corpus ofrece una herramienta con características diferentes, como son: diversas opciones de búsqueda, variados ordenamientos de las concordancias producidas, cambios en el tamaño de la ventana de la concordancia y distintos filtros de los documentos de donde se toman las concordancias, entre otros.

En el caso del CHEM, la herramienta generadora de concordancias ofrece tres posibilidades. La primera es el cambio de tamaño de la ventana de la concordancia, siendo el tamaño predefinido de 10 palabras. Muchos corpus electrónicos establecen la ventana en caracteres (letras); el mismo CHEM en su versión prototipo lo hacía de esta manera. Una ventana establecida en número de caracteres corta las palabras, aunque puede ayudar a facilitar la generación de las concordancias.

La segunda característica es el establecimiento de un filtro temporal a los documentos del CHEM. Este filtro está basado en lapsos de cien años de la siguiente manera: cincuenta años antes de la fecha que el usuario establezca y cincuenta años después. Detrás de esta idea está la intuición lingüística de que los cambios en la lengua se concluyen en por lo menos dos generaciones (hijos y nietos), aunque no se puede hablar de un parámetro fijo. Unas veces los cambios son muy rápidos y otras, muy lentos. Muchos corpus no permiten establecer el año de búsqueda y filtran por siglos (XV, XVI, XVII, etcétera); sin embargo, los cambios en la lengua no siguen estas convenciones cronológicas. La lengua está en constante evolución y las variables que influyen en sus cambios son tan variadas que pueden o no coincidir con un cambio de siglo.

La tercera y última característica del generador de concordancias del CHEM es la variedad de opciones de búsqueda. En la figura 8 se muestra la tabla que presentamos en la interfaz de consulta del CHEM con las opciones de búsqueda (para desplegar esta tabla será necesario seleccionar el enlace *Ver opciones de búsqueda*). Como ya lo mencionábamos arriba, el CHEM cuenta actualmente con tres variantes de búsqueda: por ortografía normalizada o modernizada, por coincidencia ortográfica exacta y por lema.

Tal y como se muestra en la figura 8, las búsquedas ortográficas y por lema están asociadas a ciertos códigos o símbolos. Si el usuario desea buscar las coincidencias exactas de una palabra, deberá ponerla entre comillas dobles (""). Si desea recuperar las variantes ortográficas asociadas a un lema en particular, tendrá que poner el lema entre corchetes cuadrados ([]). En seguida describiremos con ejemplos estos tipos de búsquedas.


Ver opciones de búsqueda

Tipo	Formas asociadas	Codificación	Resultados
Normalizada (modernizada)	Todas las variantes ortográficas	Ninguna, por ejemplo: vivir	vivir, bibir, bjbir
Ortográfica	Coincidencia ortográfica exacta	Doble comilla " ", por ejemplo: "bibir"	bibir
Lema	Lemas	Corchetes [ ], por ejemplo: [vivir]	vivir, viviré, vivió, vive

Para buscar subcadenas utilice *, por ejemplo: *ir, "*ido", [*ar]

Petición de búsqueda:
Año: 1550 (se obtendrá un rango de 50 años antes y 50 años después)
Ventana: 10 palabras Buscar

Figura 8. Tabla con las opciones de búsqueda disponibles en la versión en línea de la interfaz de consulta del CHEM

Como puede verse en la figura 8, las búsquedas por ortografía normalizada (primer renglón de la tabla) permiten recuperar una palabra en sus distintas variantes ortográficas, lo que salva la barrera entre la escritura de nuestros días y la de los siglos más antiguos incluidos en el CHEM. Así, ante la petición de búsqueda de la palabra «vivir» obtenemos resultados como: «vibir», «vivir», «bibir» y «bj/6vir» (véase la figura 9).


Ver opciones de búsqueda

Tipo	Formas asociadas	Codificación	Resultados
Normalizada (modernizada)	Todas las variantes ortográficas	Ninguna, por ejemplo: vivir	vivir, bibir, bjbir
Ortográfica	Coincidencia ortográfica exacta	Doble comilla " ", por ejemplo: "bibir"	bibir
Lema	Lemas	Corchetes [ ], por ejemplo: [vivir]	vivir, viviré, vivió, vive

Para buscar subcadenas utilice *, por ejemplo: *ir, "*ido", [*ar]

Petición de búsqueda: vivir
Año: 1550 (se obtendrá un rango de 50 años antes y 50 años después)
Ventana: 10 palabras Buscar

Ver estadísticas de asociación de palabras

No.		Palabra base		Referer
1	sy no comes, te morirás, qeysto es qar nos faze	vibir	", e qar no le dixo otra cosa de lo qar	1536. B
2	*ábia fecho* pedaços por ellos, y que *ábia* tomado/ a	vivir	. Dixo qar no. ¶ *Preguntado* si conoçe a *Xristoval de Çisneros*	1536. B
3	a la dicha *sentrenia*, en abjuracion pública, e juraron a	bibir	o morir debaxo de/ *nuestra* santa fe católica, so pena	1539. B
4	a su muger, qar se vengan a esta tierra a	bj /6vir	. La tierra es buena, aunqar no está como solja; pero	1572. C

Figura 9. Resultado de la búsqueda por ortografía normalizada de la palabra «vivir» en el CHEM.

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

Ahora bien, imaginemos que deseamos saber si el verbo «decir» fue escrito alguna vez como «desir» o como «dezir», y al mismo tiempo queremos recuperar las concordancias asociadas a estas variantes ortográficas. Para ello podemos auxiliarnos de la búsqueda ortográfica, que como puede verse en el segundo renglón de la tabla de la figura 10, requiere poner entre dobles comillas nuestra palabra de búsqueda. De esta manera, ponemos en la caja de petición de búsqueda “dezir” y realizamos la búsqueda. La figura 10 muestra el resultado de esta petición.

Figura 10. Resultado de la búsqueda ortográfica (por ortografía exacta) de la palabra «dezir» en el CHEM.

La tercera opción de búsqueda (tercer renglón de la tabla de la figura 11) es la búsqueda de variantes morfológicas a partir de un lema. Como ya lo explicábamos arriba, el lema representa a un conjunto de variantes. Para realizar este tipo de búsqueda debemos poner el lema entre corchetes cuadrados. Por ejemplo, pongamos en la caja de petición de búsqueda [decir] y obtengamos sus concordancias. Como puede verse en la figura 11, obtenemos «dixese», «dixeron», «dizen» y «dezían», entre otras variantes conjugadas del verbo «decir».

Figura 11. Resultado de la búsqueda por lema «decir» en el CHEM.

Finalmente, todas estas opciones de búsqueda pueden modificarse para recuperar coincidencias de subcadenas o segmentos de palabras. Esto es, imaginemos que deseamos recuperar las concordancias asociadas a los participios terminados en el segmento ~ado. Este tipo de búsqueda puede realizarse en el CHEM, poniendo el asterisco (*) como sustituto del comienzo de la palabra de petición, ya que lo que nos importa es el final de ésta. Así, y siguiendo los ejemplos que muestra el último renglón de la tabla de la figura 12, podemos realizar nuestra búsqueda como: *ado. La misma figura (figura 12) muestra el resultado. Véase como se obtienen palabras con el segmento final ~ado, como: «cavsado», «loado», «pasado», «negociado» y «determinado», entre otras.

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

Figura 12. Resultado de la búsqueda del segmento ~ado en el CHEM.

En el caso anterior hicimos una búsqueda por ortografía normalizada, ya que no le agregamos ni dobles comillas ni corchetes cuadrados al segmento buscado, pero el uso del asterisco es válido también en las otras dos opciones de búsqueda. Además, este símbolo (*) puede ponerse al principio, final o en medio de la palabra.Por ejemplo, para buscar todas las palabras que incluyen doble ese (ss), en cualquier parte de ellas podemos hacer la petición: “*ss*”, lo que traerá coincidencias ortográficas exactas que incluyan doble ese. La figura 13 muestra el resultado de esta petición.

Figura 13. Resultado de la búsqueda por ortografía exacta de palabras con doble ese (ss) en el CHEM.

En esta sección hemos mostrado las opciones de búsqueda de concordancias que brinda el CHEM. No está por demás decir que todas han sido hechas para el periodo comprendido entre 1500 y 1600, es decir cincuenta años antes de 1550 y cincuenta años después. Si se vuelve a observar las figuras de la 9 a la 13, se podrá ver que todas tienen este año como base de las búsquedas. Adelante cambiaremos este año para ejemplificar otra de las herramientas del CHEM. También es pertinente recordar que este tipo de búsquedas es posible gracias al etiquetado XML de los documentos del CHEM, descrito en secciones anteriores.

2.2.Estadísticas de asociación de palabras

Uno de los fenómenos lingüísticos que podemos encontrar comúnmente en las lenguas humanas, es la asociación de palabras. Este fenómeno tiene distintos tipos como explicaremos brevemente a continuación. Hay asociaciones que pasan a formar parte del sistema lingüístico, es decir, son tan fuertes que se usan automáticamente y son parte de la gramática de la lengua. Ejemplo de lo anterior sería la asociación entre artículos y sustantivos del español, siempre que un sustantivo aparezca con un artículo. Éste último deberá aparecer antes y concordar en género y número. No seguir esta asociación o regla nos hace transgredir la gramática de

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

nuestra lengua.

También podemos encontrar asociaciones más semánticas, donde el significado que producen las palabras asociadas es distinto a la suma de los significados de cada palabra en lo individual. Por ejemplo, si un hablante de español usa la frase hubo mano negra, seguramente quiere transmitir que alguien hizo trampa y no que apareció una mano pintada de color negro. Otras asociaciones se dan por la creación de términos compuestos, especialmente en textos especializados de cierto dominio del conocimiento humano. Ejemplos de éstos, serían: c*ampo electromagnético* o *sistema solar*. Finalmente mencionaremos las asociaciones producidas por la formalización de un nombre, como sería la *Organización Mundial de la Salud*. Existen otros tipos de asociación, pero creemos que éstas bastan para entender su importancia en el conocimiento de una lengua. Al respecto, el CHEM descubre de manera automática diversas asociaciones entre palabras.

Para determinar asociaciones entre palabras, el CHEM utiliza estadísticas de digramas. Un digrama está formado por dos palabras del corpus. Este tipo de estadísticas mide la independencia entre dos palabras y brinda una medida de dicha independencia. De entre las posibles medidas estadísticas, el CHEM utiliza la prueba de independencia de x2 (chi cuadrada), la razón de semejanza, el coeficiente de coligación de Yule y la estadística de información mutua. Ya que cada medida produce valores distintos, se obtiene un promedio normalizado con el que se determina el nivel de asociación.

El cálculo de estas estadísticas se hace tres veces y toma como datos de entrada las palabras presentes en el conjunto de concordancias generadas por una búsqueda. Es importante decir que esta herramienta trabaja de forma unida con el generador de concordancias. No está disponible de forma individual. En otras palabras, siempre que se generen concordancias para una palabra de búsqueda, se calcularán las estadísticas de asociación de las palabras de esas concordancias.

La primera vez se calcula entre la palabra base de las concordancias y las palabras inmediatamente antes. La segunda vez se calcula entre la palabra base y las palabras inmediatamente después. La tercera vez se hace entre la palabra base y todas las palabras presentes en las concordancias. El resultado de las estadísticas y el promedio normalizado se muestran en tablas ordenadas por éste último.

Veamos ahora algunos ejemplos de esta herramienta de análisis del CHEM. Busquemos en 1550 la palabra «oficio». Recordemos que la búsqueda será en cincuenta años antes y cincuenta años después. La figura 14 muestra las concordancias resultantes. Podemos observar a simple vista, y por la mera frecuencia de aparición, que existe una fuerte asociación

entre la palabra «santo» y la palabra de búsqueda. Esta asociación se debe a la existencia de una institución del siglo XVI, el *Santo Oficio de la Inquisición*. Pero veamos qué dicen las estadísticas de asociación.

Figura 14. Resultado de la búsqueda de la palabra «oficio» en el CHEM.

Para mostrar las asociaciones que propone el CHEM, será necesario acceder al enlace: V*er estadísticas de asociación de palabras*, que se puede ver en la figura 14. Cuando accedemos a este enlace aparece una ventana que muestra las tablas con las estadísticas calculadas.

La figura 15 muestra las dos primeras tablas. La primera contiene las asociaciones con el promedio normalizado más alto, entre la palabra base de la concordancia y las palabras inmediatamente anteriores. La segunda contiene las asociaciones con el promedio normalizado más alto, entre la palabra base de la concordancia y las palabras inmediatamente posteriores.

Estadísticas de palabras inmediatamente antes de **oficio** en el rango de años de 1500 a 1600

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
santo	6.5777	197.1977	12194.1092	0.9424	1

N/S = No significativo

Estadísticas de palabras inmediatamente después de **oficio** en el rango de años de 1500 a 1600

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
de	2.6252	35.1463	119.9485	0.5974	0.7486

N/S = No significativo

Figura 15. Primeras dos tablas de asociación de palabras, a partir de las concordancias de la palabra «oficio».

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

Se puede ver en la tabla de la figura 15 que las palabras más asociadas con la palabra «oficio» son «santo» a su izquierda y «de» a su derecha. Esto nos habla de la existencia de una frase muy asociada presente en el corpus: *santo oficio de*, lo que confirma la observación hecha anteriormente sobre la institución: *Santo Oficio de la Inquisición*. La figura 16 muestra la tercera tabla de asociaciones, las que se descubren entre la palabra base y el resto de palabras de las concordancias.

Estadísticas de palabras dentro de toda la concordancia en el rango de años de 1500 a 1600

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
deste	6.9556	541.7654	44010.5311	0.9651	0.9356
y	4.446	554.4177	5639.1136	0.9223	0.6421
del	5.5224	379.5084	9703.6947	0.9171	0.6151
alguacil	7.3239	146.018	16660.6606	0.964	0.6119
carcel	7.5287	124.124	16732.962	0.9696	0.6096
preso	7.2524	144.0215	15510.2397	0.9619	0.6019
santo	6.6349	211.3616	13673.0676	0.9448	0.6001
dicho	5.0124	376.9708	6400.8108	0.8982	0.5761
por	5.3974	318.8572	7456.8727	0.9064	0.5688
bienes	6.9308	135.4691	11239.3026	0.9518	0.5622
fiscal	7.3345	92.7229	10715.2642	0.9633	0.5542
la	5.0841	325.3379	5918.0318	0.8941	0.551
mayor	7.1232	101.8537	9910.6685	0.957	0.5462
en	4.8546	326.2435	4947.8023	0.8842	0.537
indio	6.2153	172.72	7978.9211	0.9285	0.5344
que	4.2886	372.9448	3576.1425	0.8695	0.5306
inquisicion	7.8651	103.537	N/S	0.9797	0.5172
lengua	6.4639	112.4745	6398.9065	0.9351	0.5069

N/S = No significativo

Figura 16. Tercera tabla de asociación de palabras a partir de las concordancias de la palabra «oficio».

En la tabla de la figura 16 podemos encontrar otro tipos de asociaciones que no forman necesariamente una secuencia de palabras (asociaciones sintagmáticas). Veamos por ejemplo que la palabra «oficio» está muy asociada con palabras, como: «alguacil», «cárcel», «preso», «fiscal» e «inquisición», entre otras. Estas asociaciones dan muestra de un campo léxico, es decir, un conjunto de palabras que se usan en un mismo contexto, que para este caso podría ser de tipo jurídico. Claro que estas asociaciones tienen que ver con los tipos de documentos, que en este caso son precisamente juicios inquisitoriales. Véase aquí la importancia de contar con documentos de diversas temáticas en el corpus.

Ahora mostraremos otro ejemplo de asociación de palabras. Para ello, hagamos una búsqueda en 1850, siglo XIX, donde el CHEM cuenta con un amplio recetario. Antes de realizar la

búsqueda de concordancias, preguntémonos con qué palabras estaría asociada la palabra «pimienta». Una vez realizado el ejercicio mental, hagamos la búsqueda. La figura 17 muestra las primeras dos tablas de asociación, que son, como recordaremos, de tipo más gramatical y sintagmático.

Estadísticas de palabras inmediatamente antes de pimienta en el rango de años de 1800 a 1900

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
y	3.13	486.9007	2189.2599	0.7106	0.8071
bastantita	6.4593	47.337	N/S	0.9556	0.5243

N/S = No significativo

Estadísticas de palabras inmediatamente después de pimienta en el rango de años de 1800 a 1900

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
gorda	6.3342	114.6084	5621.1398	0.9482	0.8346
molida	5.7197	137.8872	4244.3533	0.9116	0.7625
y	2.7578	275.7194	997.704	0.639	0.5633
despolvoreada	6.1716	43.889	N/S	0.9378	0.5118
molidos	5.0265	57.7871	1054.0887	0.8631	0.5099

N/S = No significativo

Figura 17. Primeras dos tablas de asociación de palabras, a partir de las concordancias de la palabra «pimienta».

Como puede verse en la figura 17, las palabras más asociadas a la izquierda de «pimienta» son la conjunción «y» y el cuantificador «bastantita». Esto nos habla de que tal palabra aparece comúnmente en frases que contienen una lista de elementos. Si pensamos en la palabra «pimienta» no será muy difícil pensar automáticamente en la frase: sal y pimienta. El cuantificador llama la atención, por llevar un diminutivo y nos dice que para el autor del recetario la cantidad justa de pimienta no es ni poca ni bastante, sino *bastantita*.

Sobre las palabras asociadas del lado derecho de «pimienta», éstas forman una relación adjetival. Se trata de tipos de pimienta: «gorda», «molida» o «despolvoreada». En español, el adjetivo se coloca regularmente después del sustantivo.

La figura 18 despliega la tercera tabla de asociaciones descubiertas, con base en las concordancias de la palabra «pimienta». Antes de comentar algo sobre la tabla, pensemos:

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

¿qué palabras nos vienen a la mente cuando pensamos en la palabra «pimienta»?

Estadísticas de palabras dentro de toda la concordancia en el rango de años de 1800 a 1900

Palabras	I	-2logλ	χ^2	Y	Promedio normalizado
clavo	6.6391	1741.2797	102327.0802	0.976	0.9918
canela	6.4563	1205.7742	62931.1132	0.963	0.8087
sal	5.4876	1436.6815	35098.3107	0.9229	0.7283
con	4.5941	1498.2302	18058.3261	0.8875	0.6539
un	4.8103	1300.8701	18818.9313	0.8885	0.6355
molidos	6.3765	314.5078	15837.1845	0.9522	0.56
nuez	6.2068	323.5169	14346.9064	0.9424	0.5489
poco	5.0834	718.4222	13260.0752	0.8846	0.5473
en	4.2286	1048.8554	9875.7906	0.8454	0.5453
cominos	6.0265	341.7008	13210.016	0.9323	0.5396
moscada	6.2587	270.8688	12507.8713	0.9449	0.5394
azafran	6.1325	273.4028	11478.3546	0.9376	0.5308
agengibre	6.4593	142.3336	7648.0275	0.9562	0.5193
clavos	6.2323	190.1678	8627.5696	0.9427	0.5168
gorda	6.4295	129.3668	6803.8014	0.9542	0.5138
molida	6.0251	201.9515	7831.0902	0.9307	0.5059
le	4.6928	632.6678	8711.0062	0.8566	0.5024
bastantita	6.6824	63.3259	N/S	0.9717	0.5014
tabasco	6.6824	63.3259	N/S	0.9717	0.5014

N/S = No significativo

Figura 18. Tercera tabla de asociación de palabras a partir de las concordancias de la palabra «pimienta».

La tabla de la figura 18 nos muestra cómo las estadísticas de digramas, comentadas arriba, descubren nuevamente un campo léxico, en este caso el de las *especies y condimentos*. Así, la palabra «pimienta» se relaciona fuertemente, como pudimos intuirlo al hacernos la pregunta planteada en el párrafo anterior, con otras especies, como: «clavo», «canela» y «cominos»; y con condimentos, como: «nuez», «agengibre» y por supuesto «sal», reafirmando que la frase *sal y pimienta* tiene una fuerte asociación entre sus palabras.

2.3.El acervo del CHEM

La tercera herramienta que ponemos a disposición como parte del CHEM, es lo que llamamos el *acervo*. Éste consiste en un subconjunto de documentos de la colección completa del CHEM en formato PDF y un visor que permite su lectura completa. En la figura 19 podemos ver la página de inicio del acervo, misma que es accesible mediante el menú herramientas. El acceso al acervo está disponible para cualquier persona que utiliza una computadora que forma parte

de la red UNAM y para los usuarios registrados.

Figura 19. Página de inicio del acervo del CHEM.

Actualmente el acervo cuenta con documentos de varios siglos, entre los que se incluye el recetario del siglo XIX que mencionábamos antes y cuya muestra ponemos en la figura 20. Además, contiene juicios inquisitoriales contra indígenas y cartas del siglo XVI. También se incluyen notas, peticiones y diversos documentos desde el siglo XVI hasta principios del XIX. Finalmente, se puede leer una pequeña colección de textos del siglo XVIII.

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

Figura 20. Muestra del recetario como parte del acervo del CHEM.

Es pertinente resaltar que gracias al formato XML de los documentos y su correspondiente etiquetado, nos es posible presentar los mismos con una vista muy cercana al documento original, señalando las reconstrucciones, marcas de párrafo, línea, rúbricas, notas y toda la información incluida por el editor del documento. A propósito de esto, al visualizar los documentos, en muchos de los casos desplegamos los criterios de transcripción asociados, tomados de las notas originales del transcriptor o editor, como puede verse en la figura 21.

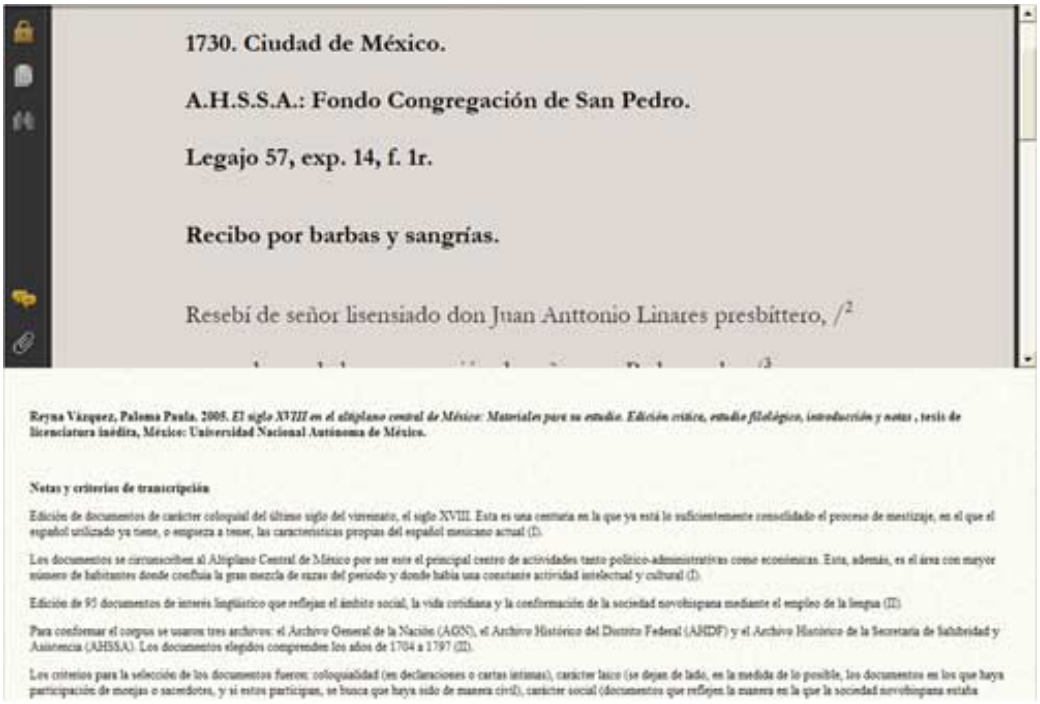


Figura 21. Muestra de los criterios de transcripción desplegados junto con los documentos del acervo del CHEM.

Finalmente, para terminar este apartado, queremos resaltar que una de las funcionalidades más interesantes del acervo es que está enlazado con las otras dos herramientas del CHEM: el generador de concordancias y las estadísticas de asociación, que ya describimos arriba. Así, cada palabra de los documentos del acervo es un enlace que permite generar sus concordancias y con ellas sus estadísticas de asociación. Los parámetros de búsqueda en este caso serán cincuenta años antes y cincuenta años después de la fecha original del documento, una ventana de diez palabras y el tipo de búsqueda normalizada.

3.Conclusiones

En este artículo hemos hecho una breve presentación del CHEM, un corpus diacrónico que documenta cómo se ha utilizado la lengua española en México desde el siglo XVI. Se trata de un corpus en crecimiento que, por ahora, cuenta con pocos documentos, en relación con otros corpus electrónicos que son mucho más vastos. En este aspecto, uno de los trabajos a futuro será compilar más documentos y prepararlos para su presentación electrónica. También mostramos las herramientas con las que cuenta el CHEM, tanto administrativas como lingüísticas. Como se vio, dentro de estas últimas se cuenta con el generador de concordancias, con la visualización de textos completos y con estadísticas de asociación. Las aplicaciones de este tipo de recurso digital son muchas y todavía están por conocerse nuevas posibilidades. Claramente, como con los otros tipos de humanidades digitales, podemos considerar algunas

1 de julio 2011 • Volumen 12 Número 7 • ISSN: 1067-6079

de ellas para investigar cómo son verdaderamente las lenguas del mundo, en este caso el español mexicano. Pero también podemos valorar otras que nos dan placer lingüístico y satisfacen nuestras curiosidades más espontáneas sobre algo tan complejo, y a la vez tan elusivo, como nuestro medio más presente, y por tanto, más invisible de comunicarnos.**4. Referencias**

Gómez, Manuel y Claudia González. 2010. *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 de licenciatura inédita. México: Universidad Nacional Autónoma de México.

Medina, Alfonso y Carlos Méndez. 2006. “Arquitectura del Corpus Histórico del Español en México (CHEM)”, en A. Hernández y J. L. Zechinelli (eds.), *Avances en la ciencia de la computación*, México: Sociedad Mexicana de Ciencia de la Computación. pp. 248-253.