UN CORPUS PARA LA INVESTIGACIÓN EN LA EXTRACCIÓN DE TÉRMINOS Y CONTEXTOS DEFINITORIOS: HACIA UN DICCIONARIO DE LAS SEXUALIDADES DESDE MÉXICO

T. A. Reyes-Careaga[*]
A. Medina Urrea[**]
G. E. Sierra Martínez[***]

RESUMEN: en este trabajo se hace una descripción de la metodología utilizada en la extracción terminológica a partir de un corpus de sexualidades en México. Se describe el proceso de compilación del corpus, así como la fase de pre-procesamiento y el etiquetado XML. Específicamente, este corpus sirve para la extracción de términos y sus contextos definitorios. En este entorno, examinamos al sistema Ecode, una herramienta de extracción de contextos definitorios. La idea es utilizar estos recursos en la generación automática de diccionarios; esto es, de listas de términos y definiciones que pueden servir de base para la construcción de diccionarios especializados. la integración propuesta de los métodos descritos constituye un recurso valioso aplicable a diversas variantes del español, en este caso, dos registros específicos: 1) del lenguaje de especialidad sobre la sexualidad y 2) del lenguaje familiar de los hablantes del español hablado en México.

PALABRAS CLAVE: Extracción terminológica, Contexto definitorio, Sexualidad, Corpus lingüístico.

RESUMO: Neste trabalho, faz-se uma descrição da metodologia utilizada na extração terminológica a partir de um corpus sobre o tema da sexualidade no México. Descreve-se o processo de compilação do corpus, assim como a fase de pré-processamento e a etiquetagem em XML. Especificamente, este corpus serve para a extração de termos e seus contextos definitórios. Neste cenário, examinamos o sistema Ecode, uma ferramenta de extração de contextos definitórios. A idéia é utilizar esses recursos para a geração automática de dicionários, isto é, de listas de termos e contextos definitórios que possam servir de base para a construção de dicionários especializados. a integração proposta dos métodos descritos constitui um recurso valioso aplicável a diversas variantes do espanhol, no caso aqui, dos registros específicos: 1) da linguagem de especialidade sobre o tema da sexualidade e 2) da linguagem familiar dos falantes do espanhol falado no México.

PALAVRAS-CHAVE: Extração terminológica; Contexto definitório; Sexualidade; Corpus linguístico.

ABSTRACT: In this paper we describe the methodology for term extraction in a Mexican corpus of sexuality. We describe the compilation process of the corpus, as well as the previous processing of texts and XML tagging. This corpus will be used for term extraction and definitional context extraction. Particularly, we examine the ECODE system, a tool for definitional context extraction. The general idea of using these resources is the automatic generation of dictionaries; namely, the automatic creation of term lists and definitions that can be used as a basis for specialized dictionaries. The integration of the methods we describe in this paper could be used as an important contribution to be applied to different Spanish registers: 1) the specialized language of sexuality, and 2) the familiar language of speakers of Mexican Spanish.

KEY WORDS: Term extraction, Definitional context, Sexuality, Corpus linguistics.

Cómo citar este artículo: Reyes-Careaga, T. A.; Medina Urrea, A.; Sierra Martínez, G. E. Un corpus para la investigación en la extracción de términos y contextos definitorios: hacia un diccionario de las sexualidades desde México. *Debate Terminológico.* No. 07, Abril 2011.; pp. 24-35

[*] Grupo de Ingeniería Lingüística, Instituto de Ingeniería, Universidad Nacional Autónoma de México. treyesc@iingen.unam.mx

[**] Grupo de Ingeniería Lingüística, Instituto de Ingeniería, Universidad Nacional Autónoma de México. amedinau@iingen.unam.mx

[***] Grupo de Ingeniería Lingüística, Instituto de Ingeniería, Universidad Nacional Autónoma de México. gsierram@iingen.unam.mx

1. INTRODUCCIÓN

Con los medios de comunicación actuales y la llamada globalización hay una renovada tendencia a la estandarización de todas las cosas, incluyendo la terminología, que es el tema que nos ocupa. Si bien a menudo se observa que la terminología de un área tiende a ser la misma en el ámbito internacional (salvo algunas variaciones), existen palabras de uso no especializado, que se refieren a los conceptos de especialidad (en este caso, de las áreas de sexualidad y sexología), que, dependiendo de la región geográfica o el estrato social, van cambiando de acuerdo con el usuario y ponen de manifiesto las características del mismo.

En el ámbito de la terminología especializada para la educación sexual en México no existe un consenso, además de que existen muy pocos glosarios o vocabularios con los términos concernientes a esta área. Por otra parte, los temas de sexualidad se han tratado durante mucho tiempo como una especie de tabú, de manera que existen palabras de uso no especializado entre la población en general que, por ejemplo, emplea diversos tipos de expresiones para disimular que se habla de determinado tema sexual y sin recurrir, obviamente, a los términos especializados[1]. En este contexto, surge la necesidad de conocer tanto los términos del área de sexualidad, como las palabras de uso no especializado, con el fin de realizar estudios sociológicos, ideológicos o para la creación de diccionarios y vocabularios para apoyar la educación sexual en México. Como es bien sabido, una manera de encontrar los términos del área de sexualidad utilizados por las autoridades del tema, así como aquellas palabras utilizadas por la población general para referirse a los mismos términos, es la utilización de un corpus de documentos con dicha información.

Esta investigación se enmarca dentro de varios proyectos del Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería (II) de la Universidad Nacional Autónoma de México (UNAM). Hasta el momento, se ha llevado a cabo la compilación del *Corpus de las Sexualidades en México*[2] y se han aprovechado herramientas ya desarrolladas en el grupo, como ECODE y DESCRIBE® con el fin de facilitar la adecuada explotación de este corpus.

En este artículo nos proponemos mostrar la manera en la que se hizo la compilación del corpus. Así, en la sección §2 examinamos las características y los criterios generales de la conformación de dicho corpus y mencionamos el pre-procesamiento que se le dio al mismo. En la sección §3 examinamos el procesamiento del corpus, el etiquetado xml que se ha llevado a cabo y la información lingüística etiquetada. En la sección §4 bosquejaremos la metodología utilizada para la extracción de términos de dicho corpus. En la sección §5

[1] Por ejemplo, en México existe el fino (a veces no tan fino) juego de palabras con connotaciones sexuales llamado "albur" que es importante para la cultura y la tradición mexicanas. En el *Diccionario de mexicanismos*, "albur*"* se define como "juego de palabras, ágil, por lo general de alusión sexual, en el cual alguien es ridiculizado", *sv*.

[2] Disponible para el público en http://www.iling.unam.mx/csmx/

presentaremos las bases de la extracción de contextos definitorios a partir del corpus. En la sección §6 se expondrán los resultados preliminares. Al ser un proyecto en curso, faltan muchas actividades por realizar, por lo que en la sección §7 se esboza el trabajo futuro y se presentan las conclusiones que hasta ahora hemos obtenido.

2. CORPUS

La necesidad de conocer los términos del área de sexualidad, tanto los especializados como los que usa la población en general, nos llevó a la tarea de recopilar un corpus lo más equilibrado posible[3]. Se necesitaba que este corpus fuera una muestra representativa de las fuentes, ya que ésta es una manera viable y práctica de conocer la terminología del área de sexualidad. Los textos fueron recopilados por varias personas, entre ellas, prestadores de servicio social interesados en erotismo o en la cultura del albur, algunos otros del área de enfermería, además de investigadores, sexólogos, psicólogos, etc. De esta manera, se aseguró que los textos contuvieran información relevante e importante del área, así como que fueran fuentes potenciales de terminología.

Los documentos que se recopilaron estaban relacionados con diversos temas de sexualidad. Se compilaron desde artículos en Google académico, artículos en revistas, foros, periódicos, *chats*, etc. También se transcribieron algunas entrevistas hechas a estudiantes de pedagogía y antropología. De esta manera, se podían tener textos con registros formales y coloquiales, pasando por registros estándares y vulgares.

Antes de proceder con el procesamiento del corpus fue necesario llevar a cabo un pre-procesamiento de todos los textos obtenidos de manera que tuvieran el mismo formato y el mismo tipo de información. Los documentos que se obtuvieron de libros o revistas en papel, por ejemplo, se escanearon y se transcribieron a través de un reconocedor óptico de caracteres (OCR), y finalmente pasaron por un proceso de verificación de los datos por un prestador de servicio social. Los documentos que se obtuvieron de internet que ya estaban digitalizados se pasaron a formato .txt para su procesamiento. Los historiales que provenían de los chats, aunque ya estaban en formato electrónico se pasaron a .txt y sufrieron un proceso de normalización, pues el lenguaje usado en los chats está plagado de "emoticones", palabras abreviadas, siglas, cambio de grafías, etc.[4]. Una vez que se realizó este pre-procesamiento, se determinó la información del documento, como la fuente, el tema, autor, etcétera.

[3] La última versión del generador de concordancias para este corpus está disponible en: http://www.iling.unam.mx/csmx/

[4] El proceso de normalización no se refleja en la versión visible de los textos de los *chats*, de manera que no se modificaron expresiones gramaticales, unidades léxicas ni grafías. La normalización se anota en un atributo de la etiqueta xml aplicada a cada palabra gráfica del corpus. De tal manera que, cuando el usuario busque en el generador de concordancias, por ejemplo, la palabra "casa", los resultados arrojados contendrán las diferentes realizaciones gráficas de las palabras que se encuentren en el corpus, como "casa" y "ksa".

El corpus cuenta actualmente con apenas 159.546 ocurrencias de palabras gráficas (tamaño del corpus en *tokens*) y 18.303 tipos de vocablos (tamaño del vocabulario del corpus, esto es, formas sin repetición). Consta de 114 archivos divididos en ocho sub-áreas, divididas a su vez en cinco niveles que contienen cuatro archivos cada uno. Las áreas temáticas consideradas hasta ahora son: fundamentos biológicos (6 documentos), respuesta y expresión sexual (11), conducta sexual (17), enfermedades de transmisión sexual (8), atracción sexual (5), educación y cultura (17) y otros sin área asignada (50).

3. PROCESAMIENTO DEL CORPUS

Para que un corpus sea verdaderamente de utilidad necesita que los documentos estén en formato electrónico y en una codificación uniforme; esto es, que se lleve a cabo la recopilación de datos contenidos en ellos que sean de utilidad al investigador. Después de la recopilación de datos propios del documento, como la fuente, el autor, la zona, el tema, etc., se obtiene también información sobre el texto en sí. Además, cada una de las palabras tiene asociada información lingüística, como la categoría gramatical y la transcripción fonológica (en algunas palabras el atributo de la normalización también contiene información, por ejemplo, en las palabras de los textos de los *chats)*. Toda esta información se agrega al documento a través de etiquetas xml. Existe un encabezado en el que se ponen los datos del documento mencionados anteriormente a través de etiquetas de elementos, atributos y valores. En el cuerpo del texto se agregan las etiquetas para cada palabra. A continuación se definen estas etiquetas:

3.1. *Clasificación de fuente*

Con esta etiqueta nos referimos al origen del documento, se informa sobre la clasificación de la fuente sin referirse a la fuente misma. Es decir, se registra de qué tipo de fuente proviene y se clasifica con un número del 0 al 5 (valores). El 0 se asigna a los documentos provenientes de Google académico, el 1 a artículos de otras fuentes, el 2 a páginas médicas, el 4 a foros, hasta el 5, que se asignará a los historiales de chat. Es decir, mientras el valor se acerque más a cero, su fuente es más formal, si se acerca más a 5 su fuente es más informal. Con esto sólo se está clasificando la fuente y no se refiere al registro ni a la coloquialidad o formalidad del documento en sí.

3.2. *Área temática*

Los documentos se clasifican manualmente de acuerdo con la temática que en ellos se plantee y se asignan a una de las siguientes áreas, mencionadas arriba:

3.2.1. Fundamentos biológicos

Se asignan a esta área aquellos documentos relacionados con alguna de las especialidades médicas, como ginecología y urología, y los de la medicina general que aborden el tema de la sexualidad, el control de la natalidad, etc. También aquellos de las ciencias relacionadas con el cuerpo humano y la función de los órganos sexuales, como la biología, la anatomía, la ginecología y la andrología.

3.2.2. Respuesta y expresión sexual

En esta área se clasifican los documentos que traten mayoritariamente cuestiones psicológicas, sociales, culturales, etc. que estén relacionadas con aspectos biológicos.

3.2.3. Comportamiento o conducta sexual

En este tema se colocan los documentos asociados a todo el entorno de la relación sexual, a la masturbación, al preámbulo de la relación sexual, a las actividades orales, coitales y anales, etcétera.

3.2.4. Identidad sexual y de género

Los documentos relacionados con los roles y las orientaciones sexuales se clasifican en esta área, así como los relacionados con los fundamentos biológicos de la identidad sexual, la identidad del sujeto, los estereotipos y las cuestiones relacionadas con género.

3.2.5. Enfermedades de transmisión sexual (ETS)

Documentos de sintomatología, análisis, control y prevención de enfermedades transmisibles sexualmente se encuentran en esta área temática.

3.2.6. Atracción sexual

Se encuentran en esta área documentos que hablan de las relaciones de pareja, desde la atracción sexual, cuestiones sociales, culturales, terapias de pareja, etc. También los documentos que tratan de acoso y abuso sexual y sus consecuencias. Las cuestiones de fidelidad se recogen en esta área.

3.2.7. Educación y cultura

Finalmente, en esta etiqueta se ubican documentos cuyo tema central son cuestiones culturales asociadas a la sexualidad en México, como la educación sexual en varios niveles (edad, sexo, ámbito social) y la educación sexual actualmente en la cultura occidental y especialmente en el ámbito mexicano.

Si el documento en cuestión no entra en ninguna de las áreas temáticas mencionadas se le asignará una etiqueta de <pendiente/>. De esta manera, si surgen nuevas áreas en los documentos del corpus, este documento podrá ser clasificado.

3.3. *Registro*

Bajo la etiqueta de <registro/> se clasifica al documento de acuerdo con la variante que se maneje en su contenido. Se pueden obtener valores desde familiar hasta culto, es decir que un documento en el que se manejen palabras de uso no especializado es clasificado como estándar, familiar, coloquial o sub-estándar; mientras que uno que contenga términos especializados se clasifica como estándar o culto.

3.4. *Fuente*

Como su nombre lo indica, esta etiqueta contiene los datos sobre el origen del documento. En el caso de artículos extraídos de Google académico, la fuente es la ficha bibliográfica de dicho documento, que contiene: autor, año de publicación, título del artículo, publicación en la que se encuentra, número de la publicación, si lo tuviere, lugar de publicación y editorial.

En el caso de publicaciones de otras fuentes electrónicas (foros*, chats*) se indica el nombre de la página, y si se tuvieran otros datos se ponen, como el nombre del autor, la fecha de la publicación y la liga a la página.

Hasta aquí hemos descrito las etiquetas con las que actualmente cuenta el corpus etiquetado. Además, incluimos otras dos etiquetas que consideramos importantes y que a continuación se describen:

3.5. *Zona geográfica*

Consideramos que es importante, en la medida de lo posible, tener el registro de la zona geográfica en la que se escribe un documento o de la que es originario el autor de dicho documento. De esta manera la información que el hablante otorgue puede resultar relevante para realizar análisis de tipo sociolingüístico o dialectal, incluso diastrático o diatópico.

3.6. *Calificación de autoridad*

Bajo esta etiqueta se califica el documento de acuerdo con la veracidad de la información que presenta. Un equipo conformado por especialistas, entre ellos sexólogos, psicólogos, antropólogos, médicos y terapeutas asignará una calificación sobre la autoridad del documento, esto sin importar su registro o fuente.

4. EXTRACCIÓN DE TÉRMINOS

Ya se ha realizado un primer acercamiento a la extracción de un vocabulario básico con nuestro corpus, usando herramientas computacionales (Lázaro, 2010). Se ha creado un vocabulario que contiene mil términos distribuidos en las ocho áreas temáticas. Es decir, se cuenta actualmente con un vocabulario de 125 términos para cada área.

En la primera etapa se generó una lista de palabras plenas, es decir, se eliminaron todas las palabras vacías que se encontraron en el corpus (conjunciones, determinantes, preposiciones). Los términos se extraen del corpus por métodos estadísticos automáticos con la ayuda de herramientas como *Wordsmith* (Sierra et al., 2009).

En una segunda etapa se creó una lista ordenada de las palabras plenas obtenidas en la etapa previa, de manera que las más frecuentes aparecieron a la cabeza de esta lista. Este proceso se hizo de manera automática pero contó también con la supervisión de lingüistas, dado que no todas las palabras que aparecieron en esta lista son términos del área de sexualidad, sino que podía haber palabras del léxico general. De esta manera se obtuvo una lista de candidatos a términos.

En la tercera etapa se procesó la lista de candidatos a términos por medio de la realización de pruebas a cada uno de los candidatos; estas pruebas se usan en recuperación de información (RI) y son el *peso* de cada palabra y el rango o *ranking* que ésta ocupa de acuerdo con su contexto, posición, situación o importancia en relación con una búsqueda.

El *peso* de cada palabra es un valor numérico asignado a ella de acuerdo con su posición en la lista de candidatos, es decir, por su frecuencia dentro del corpus general y a la frecuencia en un área temática determinada. Por ende, se determina el área a la que corresponde una palabra y el peso de la palabra de acuerdo con su ocurrencia en los textos del corpus, desde el más relevante hasta el menos relevante.

El *ranking* se mide con base en la capacidad de un sistema para separar las palabras relevantes de las que no lo son, tomando en cuenta su contexto y posición. De esta manera se determina si una palabra realmente es un término o sólo una palabra que acompaña a otra frecuentemente. Con los datos de peso y ranking se crearon las listas de términos para cada área temática de nuestro corpus y se conformó el vocabulario.

Es bien sabido que existe una gran variedad de métodos automáticos de extracción de términos. Así que, además del método expuesto, estamos experimentando con otros menos supervisados; específicamente el de valores C/NC de Frantzi (2000) o el YATE que combina métodos híbridos y emplea *EuroWordNet* para validar sus candidatos (Vivaldi, 2001).

5. EXTRACCIÓN EN CONTEXTOS DEFINITORIOS

Existen diversas maneras de identificar contextos definitorios en un corpus. Una vez hecha la extracción de términos, se pueden buscar ciertos patrones típicamente utilizados en la presentación de definiciones, por ejemplo los términos acompañados de estos patrones, como los tipográficos que, por cuestiones de énfasis, los autores suelen utilizar como recursos visuales para enfatizar la introducción de sus definiciones.

Esencialmente, este procedimiento se lleva a cabo con el sistema ECODE (Alarcón et al., 2008)[5], que es un sistema desarrollado para la extracción automática de información útil en la definición de los términos y que llamaremos aquí contextos definitorios (CD).

El sistema está compuesto por tres módulos relacionados con: 1) la extracción de ocurrencias de patrones verbales definitorios (por ejemplo, "se define como", "sirve para", etc.); 2) la eliminación de contextos no definitorios; y 3) la identificación en el contexto de los elementos constitutivos pertinentes (el término, el patrón –verbal o tipográfico- y el candidato a definición).

Así, con la herramienta ECODE se obtiene un análisis automático de un corpus que contiene una gran cantidad de información que, una vez extraída, sirve para construir un banco de datos de información léxica. De hecho, esta base de datos puede verse como una primera aproximación a un glosario de los términos extraídos.

Para ilustrar el procedimiento, se listan a continuación algunos ejemplos de contextos definitorios extraídos del corpus:

1. …<t>estenosis del meato</t>, que <pv>se define como</pv> <d>obstrucción o estrechez del orificio del pene por donde se expulsan orina y semen</d>

2. <d>La atención excesiva a ciertas partes del cuerpo femenino (mamas, nalgas, piernas)</d> <pv>se le conoce como</pv> <t>parcialismo</t>

3. El <t>masoquismo</t> <pv>se define como</pv> <d>la obtención de placer sexual que obtiene un individuo al ser dañado físicamente, amenazado o sometido a distintos tipos de abusos</d>

4. La <t>violencia de género</t> (que en la mayoría de los casos se dirige hacia las mujeres) <pv>se define como</pv> <d>"Todo acto de violencia por el género de la persona, que tenga o pueda tener como resultado un daño o sufrimiento físico, sexual o psicológico, incluso las amenazas de tales

[5] Una presentación de los resultados de la herramienta, en el contexto de investigación del vocabulario del genoma humano, puede verse en http://brangaene.upf.es/ecode/.

actos, la coacción o la privación arbitrara de la libertad tanto si se produce en la vida pública como en la privada" (ONU, 1993) </d>

5. <t>sistema inmunológico</t>: <pv>es</pv> <d>el sistema del organismo que sirve para protegerse de los virus y de las bacterias, así como de otras substancias "extrañas" al mismo</d>

Como se puede ver, los términos están marcados con la etiqueta <t/>, los patrones verbales con <pv/> y los candidatos a definiciones con <d/>. En estos ejemplos se refleja el hecho de que el patrón verbal "se define como" es típico de estos contextos. También se muestran patrones con otros verbos, *como conocer y ser.*

La posibilidad de extraer este tipo de información de un corpus marcado con datos que reflejan si sus documentos son especializados, si son más o menos confiables académicamente, o si son de carácter coloquial, semi-estándar, etc., permite clasificar los contextos definitorios según estos rasgos y generar semiautomáticamente diversos tipos de diccionarios y glosarios para diferentes audiencias y objetivos.

6. RESULTADOS PRELIMINARES

Describimos en el apartado §4 de este artículo la metodología con la que se extrajeron los primeros términos para la conformación de la *Terminología Básica de las Sexualidades en México*. En esa primera fase se obtuvieron 1.285 términos, de los cuales se seleccionaron 125 por cada área para sumar un total de mil términos. Asimismo, en la sección §5 se mostraron ejemplos de la extracción de algunos contextos definitorios.

Actualmente varios estudiantes de licenciatura, maestría y prestadores de servicio social han realizado extracciones de términos y de definiciones con el propósito de mejorar el sistema DESCRIBE® o de mejorar la base de datos para optimizar un motor de búsqueda para un diccionario onomasiológico de sexualidad. En total se han obtenido aproximadamente 27 mil definiciones para 291 términos; sin embargo, aún se realizan revisiones manuales para descartar aquellos candidatos a definiciones que no lo son. Es decir, como era de esperarse, el sistema obtiene buenos y malos candidatos a definiciones y los resultados deben depurarse con el fin de conocer los errores y posteriormente poder realizar mejorías al sistema.

Explicaremos lo anterior con ejemplos de buenos y malos candidatos a definiciones para tres términos: "diafragma", "travestismo" y "transexualidad".

Término	Buen candidato	Mal candidato
diafragma	el <t>diafragma</t> <pv>consiste en</pv> <d>un capuchón de goma flexible que se introduce en la vagina de forma que quede cubierto el cuello del útero</d>	el diafragma debe ser usado en combinación con un espermicida o crema
travestismo	el <t>travestismo</t> <pv>es</pv> <d>el deseo de un cierto grupo de hombres de vestirse como mujeres o de mujeres de vestirse como hombres</d>	el travestismo no es una condición propia del hombre homosexual, existen hombres heterosexuales transvestís
transexualidad	la <t>transexualidad</t> <pv>se puede definir como</pv> <d>una seguridad interior duradera de pertenecer al otro sexo</d>	este psiquiatra describe la transexualidad como el desarrollo "no normal" de la identidad sexual, en el sentido de que se sale de la estadística mayoritaria, con una prevalencia de uno por mil habitantes

Tabla 1.Buenos y malos candidatos a definiciones de: “diafragma”, “travestismo” y “transexualidad”.

En la tabla anterior mostramos, en la primera columna, los términos seleccionados; en la segunda columna, los buenos candidatos a definiciones; y en la tercera, los malos candidatos. Los buenos y malos candidatos, como se mencionó, se han distinguido manualmente, para identificar los errores del sistema. Los buenos se seleccionan si cumplen con las características que debe tener un contexto definitorio (tienen un término, <t/>, un patrón verbal <pv/> y una definición <d/> –marcados con sus respectivas etiquetas en los ejemplos- y no se atiende a la veracidad o la exactitud de la definición. Por otro lado, observamos en la tercera columna que los malos candidatos suelen contener el término, pero no poseen ya sea un patrón verbal definitorio, como el descrito aquí, o la definición del término.

7. CONCLUSIONES Y TRABAJO FUTURO

Como se ha estado mencionando a lo largo de este escrito, una de las utilidades de la creación del corpus de las sexualidades en México será hacer estudios sociológicos, dialectológicos (diastráticos y diatópicos) e inclusive ideológicos (por informantes según su sexo, edad, origen, etc.) que pudieran ser útiles para mejorar la educación sexual en México.

De la misma manera, la extracción de términos especializados y palabras de uso no especializado se puede utilizar para la creación de diccionarios y glosarios de la terminología del área de sexualidad, de forma que las definiciones de los términos estuvieran al alcance de una mayor cantidad de población. Con este acceso a la información se podrían prevenir ETS, embarazos no deseados, violencia sexual, etcétera.

Por otra parte, la terminología extraída serviría como vínculo entre la población y los médicos o servidores públicos de diferentes sectores, como médicos, psicólogos, trabajadores sociales, sexólogos, educadores, entre otros.

Además, como parte de la conformación del corpus, aún falta identificar la información faltante y necesaria, ubicar más textos, pre-procesarlos (transcribirlos, digitalizarlos y/o unificarlos en su formato) y agregarlos al corpus. Para la parte del etiquetado, falta investigar la necesidad de etiquetas adicionales para la conformación del corpus. Respecto de la etiqueta de <calificación de autoridad/> todavía es necesario que los especialistas evalúen los documentos y asignen esta calificación. Para ello será importante que estos especialistas realicen acuerdos sobre la manera de evaluar cada documento y las calificaciones que serán asignadas.

En el caso de las áreas temáticas se debe ampliar el repertorio con el fin de distribuir los documentos que actualmente se encuentran etiquetados como pendientes. Asimismo, se deben explorar métodos para que se puedan asignar documentos a varias áreas temáticas, de una manera documentada, pues es posible que un mismo documento contenga varios puntos de vista o que hable de diferentes temas.

Como parte del trabajo terminológico y lexicográfico, el trabajo futuro significa separar los términos especializados y las palabras de uso no especializado que se usan actualmente en México. La información extraída puede servir como punto de partida para la creación de diversos tipos de diccionarios y glosarios, especializados y de lengua general.

La integración que proponemos de los métodos mencionados arriba constituye un recurso muy valioso aplicable a diversas variantes de la lengua española; en el caso de este trabajo, dos registros específicos: 1) del lenguaje de especialidad sobre la sexualidad y 2) del lenguaje familiar de los hablantes del español hablado en México. Ambos registros constituyen un reto importante aún no cubierto, ya que el corpus todavía necesita crecer más en ambos sentidos. Lo importante es que se pueden extraer semiautomáticamente una amplia variedad de terminologías.

AGRADECIMIENTOS

Este trabajo ha sido posible gracias al apoyo económico del CONACyT mediante el proyecto 105711, “Extracción de conocimiento lexicográfico a partir de textos de internet”.

BIBLIOGRAFÍA

Alarcón, R., G. Sierra y C. Bach. 2008. “ECODE: A Pattern Based Approach for Definitional Knowledge Extraction”. Barcelona: XIII Congreso Internacional EURALEX.

Diccionario de Mexicanismos. 2010. México: Academia Mexicana de la Lengua.

Frantzi, K., S. Ananiadou y H. Mima. 2000. “Automatic recognition of multi-word terms: the C-value/NC-value method”. International Journal on Digital Libraries, 3:2, 115-130.

Lázaro, J. 2010. *Extracción de la terminología básica de las sexualidades en México a partir de un corpus lingüístico*. Tesis de licenciatura inédita. México: Universidad Nacional Autónoma de México.

Medina, A. 2001. "Diccionario de terminología de la sexualidad en México", en *IV Congreso Nacional de Educación Sexual y Sexología*. México: Federación Mexicana de Educación Sexual y Sexología.

Medina, A. 2003. "Construcción de un sistema lexicográfico para la educación sexual", en *V Congreso Nacional de Educación Sexual y Sexología*. México: Federación Mexicana de Educación Sexual y Sexología.

Medina, A. 2006. "Sexualidad, sexología y terminología: Hacia un corpus de las sexualidades en México", en *VI Congreso Nacional de Educación Sexual y Sexología*. México: Federación Mexicana de Educación Sexual y Sexología.

Medina, A y Sierra, G. 2004. "Criteria for the Construction of a Corpus for a Mexican Spanish Dictionary of Sexuality", en *Proceedings of the 11th EURALEX International Congress*. Lorient: Université de Bretagne Sud.

Sierra, G., A. Medina y J. Lázaro. 2009. "Determinación de la terminología básica en sexualidad a partir de la web como corpus", en *Actas del 1er congreso Internacional de Lingüística de Corpus.* Murcia: AELINCO.

Sierra, G., A. Medina y J. Lázaro. 2010. "Terminótica y sexualidad: un proyecto integral", en *XII Simposio Iberoamericano de Terminología RiTerm 2010*. Buenos Aires.

Vivaldi, J. 2001. "Extracción de candidatos a término mediante combinación de estrategias heterogéneas". Tesis de doctorado. Universitat Politècnica de Catalunya.