UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

COLEGIO DE LETRAS HISPÁNICAS

EXTRACCIÓN DE LA TERMINOLOGÍA BÁSICA DE LAS SEXUALIDADES EN MÉXICO A PARTIR DE UN CORPUS LINGÜÍSTICO

T E S I S

QUE, PARA OBTENER EL TÍTULO DE

LICENCIADO EN LENGUA Y LITERATURAS HISPÁNICAS,

PRESENTA:

JORGE ADRIÁN LÁZARO HERNÁNDEZ

ASESOR: DR. GERARDO EUGENIO SIERRA MARTÍNEZ

CIUDAD UNIVERSITARIA, 2010

A mi abuela. Deuda saldada.

"Mire parcero: no somos nada. Somos una pesadilla de Dios, que es loco"

Fernando Vallejo.

Agradecimientos

Quiero agradecer a todas las personas y entes que han estado a lo largo de este terrible, tortuoso e inminente proceso del paso a la vida adulta de a de veras.

En primer lugar deseo agradecer a Dios, ese otro Yo que cada uno llevamos dentro y que nos da el impulso y el "repulso" por vivir cada día. Esa ancla de la vida que nos lleva, cuando se ha roto la cadena que nos impide volar, a lo más profundo del cielo, allá donde nos achicharramos junto a la moral diaria, el ímpetu famoso, la gloria ensalzada; allá muy cerca del sol, donde se encuentra el verdadero infierno: uno mismo. Gracias por estar acribillándome a diario.

A mis abuelos. Por criarme y mantenerme, por hacerme crecer. A mi abuela por su insistente, pertinaz y agudo regaño, por enseñarme a ser congruente conmigo mismo, por la disciplina férrea, por la terquedad que engendré en sus brazos desde que nací. Por sembrar en mí el semblante duro, el paso firme y el interior cálido que toda persona debería tener. Por ser madre. A mi abuelo por mostrarme el mundo sin penas, con la sonrisa en la boca. Por la comprensión y la complicidad. Porque has sido siempre mi único y verdadero padre, porque gran parte de lo que le agrada a otras personas, y si no he sucumbido a la locura o amargura total, te lo debo a ti. Por la esperanza de que mañana será un día mejor, aunque no siempre coincidamos.

A mi asesor, el Dr. Gerardo Sierra. Porque te puedo hablar de tú. Gracias por ser amigo, confidente, y guía académico. Porque no me dejaste fuera y no has permitido que me salga del GIL. Por cada uno de los escollos que hemos superado juntos y por aguantarme aun cuando te he dado más problemas que tus dos hijos juntos. Gracias por introducirme en este mundo de la lingüística computacional.

A mis sinodales. La Mtra. Margarita Palacios, porque nos tratas como hijos y nos exiges como alumnos, por la voz dulce y el comentario acertado; por ayudarme con la burocracia del proceso de titulación. Al Dr. Alfonso Medina, semillero del proyecto del cual se desprende este trabajo, exigente hasta las muelas y académico de corazón. A la Dra. Adriana Ávila, por su disposición y su apoyo, porque nunca dijo que no mientras sabía que se podía continuar. A la Dra. Iria da Cunha, porque rehiciste la tesis aunque

esta no sea la versión final; si algo me hacía falta era ese rigor científico y el ojo en el detalle.

A mi tía Jessica porque somos iguales en polos distintos. Por nunca desistir en el camino que todos encuentran contrario, por desafiar lo predicho, por alzarle las faldas a la mudas. Junto a ti he disfrutado de hacer del mundo un cacahuate.

A mi madre, porque hiciste lo que debiste y pudiste en el momento exacto. Porque me diste una infancia cuasinormal que me acerco al verdadero mundo, a la realidad sin ropa, a la cruda realidad insulsa. Porque tú dejaste sobre la mesa el primer libro que leí y no has soltado nunca el último acento en nuestras vidas.

A Ivonne Padilla, única e indescriptible. A ti hacen reverencia la bondad, la paciencia, el amor y todo lo que soy dentro y fuera de mí. Porque me has sabido aguantar y porque siempre estás a mi lado más que yo mismo. Porque siempre has sido aquella a quien nombro cuando he estado más solo, más enamorado, más triste, más iluminado. Porque eres el motivo por el cual esto se escribió con el fin de continuar a tu lado. Por ser de mí y para ti.

A la familia De la Peña-Rodríguez, en especial a mi hermano y amigo, Jesús De la Peña. Porque hemos luchado juntos y porque hemos descubierto que debajo del lodo a veces hay tesoros. Por apoyarme en todo lo que hago aunque parezca lo peor para el mundo. Por las incontables historias y los muchos recuerdos juntos. Por el "Dúodeno", las letras, los cantos, las lagrimas y las heridas. Porque nunca me has dejado solo y porque lo has hecho cuando te lo pido. A má Tere, por adoptarme por temporadas enteras y por apoyarme desde mi infancia. Por grabarme en la frente el sello imborrable de la premisa de mi destrucción. Al Dr. Jorge De la Peña, por ayudarme cada vez que lo he molestado y por soltarme aquella frase que nunca olvidaré, ambos sabemos cuál y cuándo.

A todo el Grupo de Ingeniería Lingüística de la UNAM. La banda, mi Banda. Muy especialmente a Alejandro Rosas, amigo de corazón, compañero de penas (y también hágase el cambio de la nasal por la dental), colega y mago de las soluciones efectivas y rápidas a problemas inconmensurables. A Irasema y Victor, por acompañarme

siempre, adonde fuera, por aguantarme y aguantarse a mi lado. Por supuesto al H. Dr. Rodrigo Alarcón por mostrarme el camino de la rectitud en los actos académicos y la camaradería fuera de ella; porque me mostraste lo difícil que es ser una buena persona en estos días. A Josué, José Luis y Pavel, la pandilla ingenieríl, grandes personas y excelentes seres humanos. A Octavio, Teresita, Ariadna, Claudia, Jessi, Brenda, Adriana, Alicia (Alisa!!) y Carlos, porque hay algo de ustedes en estas páginas. Gracias. GIL, baby!

A Adriana Ballesteros, la única amiga que llora por mí cuando me voy (sea a Europa o al mercado). Por los domingos fatídicos y por todos estos años, porque contigo he vivido todo lo que no he querido. Por ser LA amiga. A Esther Alvarado, poetisa perdida, escritora de niños, amiga con alas. Gracias por siempre creer en mí, gracias por el café, las kilométricas charlas y todo, todo lo que me diste y nunca me cobraste. Seguirás siempre conmigo aunque no nos veamos. Gracias amiga.

A mi Mac, porque en ella vacié este trabajo y nunca me falló. Por ser la amiga que siempre cargaba en mi espalda.

Este trabajo se llevó a cabo gracias a los apoyos de PAPIIT: *Extracción de relaciones semánticas a partir de definiciones en textos de especialidad*, registro 403108 y de CONACyT: *Extracción de relaciones léxicas para dominios restringidos a partir de contextos definitorios en español*, registro 82050 y *Metodología para la extracción de términos de un área de especialidad a partir de un corpus lingüístico*, registro 101895.

1. Introducción

1.1. Antecedentes e interés del tema

El Grupo de Ingeniería Lingüística (GIL) se ha dedicado desde hace una década a la resolución de problemas lingüísticos a través de la utilización de recursos computacionales, esto con el fin de proporcionar herramientas confiables al estudioso en las áreas de lingüística que reduzcan considerablemente el tiempo de inversión en la investigación y mejoren la obtención de resultados al tratar grandes cantidades de información.

Una de las principales líneas de investigación dentro de este grupo es la lexicografía. Ahora bien, por ser un grupo interdisciplinario y con el afán de crear nuevos caminos en la investigación lingüística se ha desarrollado un proyecto que persigue la creación y distribución de un *Diccionario de la Sexualidad en México*, cuya gestación se da en el ámbito de la lexicografía apoyada por recursos computacionales.

El primer acercamiento a este proyecto lexicográfico se dio con la presentación del artículo *Criteria for the Construction of a Corpus for a Mexican Spanish Dictionary of Sexuality* publicado por Medina y Sierra (2004) en las actas del *11th Euralex International Congress*.

Posteriormente con el apoyo de CONACyT para *Proyectos de apoyo para investigadores nacionales para el fortalecimiento de actividades de tutoría y asesoría de estudiantes de nivel licenciatura (SNI-estudiantes)* se diseñó un proyecto denominado *Metodología para la extracción de términos de un área de especialidad a partir de un corpus lingüístico* con número de registro 101895 cuyo responsable fue el Dr. Gerardo Sierra en corresponsabilidad con Jorge Adrián Lázaro Hernández, el cual tenía como finalidad mostrar justamente el punto central de la presente tesis: la metodología para la extracción de términos de manera semiautomática.

Más tarde, este último proyecto dio paso a la creación de un proyecto de Ciencia Básica 2008 también de CONACyT llamado *Extracción de conocimiento lexicográfico a partir de textos en Internet* número de registro 105711 bajo la responsabilidad del Dr. Alfonso Medina. Este proyecto, ya ampliado, persigue no sólo la extracción

terminológica del CSMX sino también la creación automática de las definiciones de dichos términos con la ayuda del programa DESCRIBE®.

Así, se ve que la creación de un *Diccionario Básico de las Sexualidades en México* es de suma importancia y tiene ya antecedentes que nos permiten mostrarlo como un producto viable y necesario tanto para la investigación lexicográfica y terminológica como para la comunidad en general en nuestro país.

1.2. Planteamiento del problema

Abordar el tema de la sexualidad humana es un verdadero reto. Encontramos, en principio, una dispersión conceptual aunada a una gran cantidad de elementos sociales y culturales que han frenado la construcción de herramientas que muestren el conocimiento que tenemos acerca del tema que ahora abordamos, al menos en países como México. Por esto, es indispensable contar con recursos que faciliten el consenso, la integración y la claridad de conceptos y definiciones que permitan un entendimiento común de la sexualidad. Debido a la barrera de comunicación existente entre especialistas, y entre éstos y la población en general, es necesaria la creación de recursos que contemplen una terminología armonizada y sistemática. Si bien existen glosarios y vocabularios sobre el tema, la mayoría de ellos son dispersos. Por lo anterior, hace falta una recopilación estructurada con criterios homogéneos sobre la información de esta área de conocimiento que responda a los retos actuales que nos supone la construcción del conocimiento en sexología y el impacto del ejercicio de la sexualidad en México.

1.3. Hipótesis

Si atendemos al criterio de Teresa Cabré para referirse al lenguaje especializado como los "subcódigos que usan los hablantes y que seleccionan a tenor de las necesidades expresivas y las características particulares del contexto comunicativo" (Cabré, 2004d), caeremos en la cuenta de que hay una necesidad intrínseca del individuo por comunicarse con un grupo que comparte conocimiento especializado, para asegurar la realización efectiva de algunas tareas especializadas.

Mucho se ha dicho sobre este procedimiento de comunicación y sobre los mecanismos que lo rigen, pero poco se ha hecho por formalizarlo en un método automático o semiautomático aplicable a cualquier área. Partiendo de esto, en el presente proyecto se propone crear una metodología aplicable a cualquier área temática a través del estudio de los procesos que se llevan a cabo para la extracción de información. Es decir, si podemos desarrollar algún mecanismo de adquisición de léxico sobre un área, definir su conformación en bloques (que permitan la utilización del léxico especializado en los contextos adecuados) e implementar la selección y agrupación de conceptos, tendremos una metodología unificada para el trabajo terminológico y lexicográfico.

Aunque existen desde hace varios años una serie de investigaciones y desarrollos –extractores automáticos- para el trabajo terminográfico, nos hemos avocado en diseñar y describir esta metodología con el fin de mostrar la efectividad de esta en el desarrollo de la creación de productos eficientes y confiables.

Trabajos recientes en procesamiento del lenguaje natural han propuesto una metodología para la construcción de *Sistemas de Organización del Conocimiento a partir de un corpus documental* (Sánchez-Cuadrado et. al 2007), que es lo que más se acerca a la metodología para la extracción de términos. Por su parte, el *Tutorial Pavel de Terminología* (Pavel et. al. 2002) dedica sólo una parte de todo su material a la extracción terminológica. La Dra. Teresa Cabré (1998) tiene un amplio e interesante artículo sobre los distintos métodos de extracción terminológica aplicables en lengua española, pero no hay mención de alguno mexicano. Por lo anterior, podría decirse que casi nadie, en México, ha propuesto un trabajo dedicado exclusivamente a la metodología para extracción de términos; es aquí donde radica la importancia del trabajo actual.

Ahora bien, nuestra hipótesis plantea que esta metodología será capaz de llevarnos hasta la obtención de una terminología básica en un área de especialidad a través de la utilización de teorías y recursos de índole interdisciplinar como la lexicografía computacional y la terminótica; la segunda de ellas es muy cercana y, de hecho, parte de la primera. Este trabajo es la fusión, principalmente, entre una teoría terminológica clásica con base en los trabajos sobre todo de la Dra. Teresa Cabré y los

recursos computacionales para la obtención de vocabularios y definiciones que se han desarrollado y utilizado en el Grupo de Ingeniería Lingüística de la UNAM.

1.4. Objetivos

Para llevar a buen puerto el proyecto lexicográfico del Grupo de Ingeniería Lingüística se han planteado varias etapas que van desde la conformación de un corpus hasta la extracción automática de definiciones de los términos obtenidos. Estas etapas desembocarán en objetivos particulares que en conjunto armarán la terminología básica que dará pie al *Diccionario Básico de las Sexualidades en México*.

1.4.1. Objetivo general

El objetivo general del presente trabajo es proponer una metodología de trabajo para la extracción de términos en el área de la sexualidad, puesto que dicha área es un terreno poco estudiado por las tecnologías del lenguaje en México. A través de un corpus de textos extraído de Internet, se pretende explicar el proceso lingüístico-computacional que se sigue para la extracción de una terminología básica (Sierra, Medina, & Lázaro, 2009). Se trata de un proyecto orientado a la lexicografía computacional basada en corpus lingüísticos que intenta resolver problemas en el proceso de selección de los términos de un área en específico.

1.4.2. Objetivos particulares

- Creación de una nueva versión, ampliada y corregida, del *Corpus de las Sexualidades en México* (CSMX), que actualmente se encuentra albergado para su consulta pública en http://www.iling.unam.mx/csmx.
- Diseño y creación de una *Terminología Básica de las Sexualidades en México* cuyo cuerpo contendrá los mil términos más representativos obtenidos a partir del CSMX.
- Aportar parte del material indispensable para la definición de los términos encontrados (el corpus).

1.5. Organización de la tesis

En el capítulo 2 se presentará el estado actual de la investigación lingüística basada en corpus y de la lingüística de corpus como una disciplina necesaria para la investigación en general, además de describir cómo se ha abordado desde la lingüística teórica y desde la lingüística computacional. Como punto clave, el capítulo girará en torno a las ventajas de contar con un corpus de trabajo en soporte electrónico y los requerimientos que exige una labor como ésta, es decir, cuáles son los criterios que se deben acatar con el fin de que el conjunto de textos tenga el impacto deseado (que sea un corpus bien formado) y que al mismo tiempo responda a las necesidades que exige un proyecto lexicográfico como el que ahora se trata de desarrollar: que refleje el dialecto mexicano en sus variantes y con un buen índice de representatividad con el fin de que los términos extraídos sean realmente los que usa el hablante de este país.

En el capítulo 3 se hablará sobre el proceso de recopilación y administración de los textos que conforman el corpus de trabajo. Se abordarán temas como la recuperación de los textos desde la Web, la fase de limpiado de dichos textos y la creación de la base de datos que alberga todo lo relacionado con la administración de los documentos. Además, también se hablará sobre la fase de etiquetado, que es uno de los procesos más detallados debido a la necesidad de estandarización de las entidades de marcaje en los corpus lingüísticos actuales con el fin de hacerlos accesibles a la mayoría de los interesados en el tema; para esto, es primordial que se explique cuáles herramientas fueron utilizadas (XML y TOK) ya que cada una de ellas está basada en criterios computacionales y lingüísticos definidos. Adicionalmente, se hablará sobre el proceso del diseño de la interfaz –hecha previamente por otras personas- y la reorganización de esta para hacerla más ergonómica y amigable al usuario. Es importante este último punto, ya que dicha interfaz será finalmente lo que el consultante verá y utilizará para la explotación de este corpus desde la comodidad de su casa y/o sede de trabajo (es decir, una interfaz como la del CORDE o la del CREA).

El capítulo 4 describirá las herramientas utilizadas para el procesamiento del corpus ya diseñado para describir cuáles son los recursos que utilizó el Grupo de Ingeniería Lingüística con el fin de hacer una herramienta eficiente tanto en la parte

teórica que abarca la realidad lingüística de México en esta área, como técnica, la rapidez en el preprocesamiento con la herramienta TOK y el mejor lenguaje de marcado, XML, para la visualización de la interfaz del primer desarrollo que surge de este proyecto: el buscador de concordancias.

El capítulo 5 contendrá, de entrada, el bagaje teórico necesario para sustentar la obtención de la terminología básica de la cual se habla en el título. Tiene como primer apartado una distinción entre terminología y terminótica, debido a que la primera, salvo contados casos –y cabe mencionar muy respetables y afortunados- no se ha estudiado en México con la intensidad que se ha hecho en otros países; y la segunda como una subdisciplina emergente que justamente integra el conocimiento teórico y el uso concienzudo de las herramientas computacionales para llevar a cabo la labor terminológica. Más tarde se explicarán los procesos actuales que atañen a la extracción terminológica desde el trabajo realizado en el GIL y la metodología que hemos utilizado para obtener la terminología básica que se presentará como cumbre de este trabajo.

Finalmente, en el capítulo 6, *conclusiones*, nos concentraremos en demostrar el uso inminente de las computadoras en pro de la lingüística y cómo éstas ayudan en la conformación de materiales de trabajo o resultados más finos en la investigación. Nos centraremos, como broche, en explicar las ventajas del método descrito en la persecución del producto final inmediato al que va destinado esta investigación (la creación de la *Terminología Básica de las Sexualidades en México*) y la aportación del sistema de extracción automático de definiciones: DESCRIBE®, el cual se desarrolla en el Grupo de Ingeniería Lingüística, como trabajo a futuro.

2. El corpus lingüístico y la lingüística de corpus

Existe un asunto indiscutible en los estudios lingüísticos actuales: la gran cantidad de información que debe manejar el investigador es, la mayoría de las veces, una tarea que requiere el uso de herramientas provenientes del avance tecnológico; particularmente de la computadora.

Estos estudios están sujetos a los datos que se conservan en diferentes formatos electrónicos desde el *boom* de la Web y las redes de trabajo, que atienden por un lado a la practicidad de su consulta y, por otro, a la rapidez en cuanto al procesamiento se refiere. Es decir, la computadora no sólo ha venido a ser una herramienta indispensable en la vida cotidiana, sino que ha permeado hasta en las tareas más especializadas llevadas a cabo por un sinfín de personas en las diversas áreas del conocimiento.

Ahora, la lingüística por ser un área que trabaja con lenguaje natural es uno de los campos que más se ha apoyado en los nuevos recursos con el fin de agilizar y hacer más fina su tarea, ya que cualquier lengua, por pequeña que sea en su inventario –fónico o léxico, por nombrar alguno-, es tan vasta como para salirse de las manos de cualquier procesamiento manual. De tal manera que los corpus lingüísticos se han convertido en el bagaje primordial para casi todos los estudios que se realizan en la actualidad y son, por tanto, la materia prima de la cual depende el posible carácter exhaustivo del trabajo lingüístico.

2.1. La lingüística de corpus y su desarrollo en México

En este trabajo tomaremos como punto de partida la definición de Sierra (2006) para explicar lo que es un corpus lingüístico con el fin de no ahondar en la multiplicidad de definiciones que hay sobre este término ya que, por variadas y distintas, abordan desde puntos de vista particulares el meollo del asunto. Sierra nos dice que un corpus textual es *"la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos"*. Con lo cual caemos en la cuenta de que pueden existir colecciones de textos sobre una lengua determinada –escrita o hablada- que sin embargo, por no cumplir con criterios específicos de recopilación, se ven destinadas a no ser más que eso: una recopilación vacía que no lleva en sí un fin primordial.

Hay que mencionar, empero, que algunos de estos textos no fueron concebidos como corpus sino que fue a través de su explotación y estudio que se convirtieron en lo que hoy son. Así, podemos nombrar por ejemplo el conjunto de glosas que documentan los cambios que sufría el latín hacia el siglo XI (las cuales con una clasificación más minuciosa son llamadas ahora Silenses y Emilianenses) y las colecciones de textos, sobre todo literarios, que utilizaban los gramáticos del XIX para ejemplificar los fenómenos lingüísticos de la época.

Los primeros corpus electrónicos de los que se tiene razón datan de los años 60 y estaban codificados en tarjetas perforadas que sólo podían leer las computadoras (Procházková, 2006). Antes de esta época, es decir, de los años 50 hacia atrás, se puede concebir una especie de bagaje teórico que podríamos denominar como lingüística de corpus, debido al trabajo de personas como Kaeding, quien hacia 1897 intentaba estudiar la distribución de las frecuencias de las letras y su secuencia; Thorndike en 1921 y M. West en 1953 se proponían seleccionar las palabras más frecuentes con fines didácticos(Sánchez, 2008) y ya más adelante Boas (en su célebre *Race, Language and Culture* de 1940) proponía que el lenguaje es parte fundamental de la verdad de una cultura y, por lo tanto, debía ser resguardado de alguna manera con el fin de tener datos que sugirieran cómo esta actúo en su época. Ellos y muchos otros más ya estaban trabajando sobre estudios que permitieran definir la metodología de trabajo no de una teoría lingüística en particular, sino sobre esta nueva forma de trabajo o herramienta; estos autores han sido llamados por algunos como los teóricos de la "etapa joven de la lingüística de corpus" (McEnery & Wilson, 1996).

Como podemos imaginar, la llegada de Chomsky al terreno de la lingüística teórica dio un giro interesante a estos estudios y modificó los puntos de vista que se tenían acerca de los trabajos futuros. Entre otras cosas, estos teóricos apostaron por el camino empirista[1] para demostrar los fenómenos de la lengua en oposición a Chomsky, como nos dice Sánchez (2008, pág. 3):

[1] Entre otras corrientes. Podemos contar también, como nos muestra Caravedo, R. (1999), con una perspectiva estructuralista de tipo descriptivo para el análisis de todas las lenguas incluyendo las ágrafas que pueden diferir bastante de la postura chomskiana.

Si la teoría generativista de Chomsky había centrado la atención en la formulación de una teoría de base cognitiva e hipotético-deductiva, cuya finalidad es explicar racionalmente cómo la mente humana genera y procesa el lenguaje, la perspectiva empírica se centra en la observación del *output* lingüístico para determinar lo que es normal en la lengua.

Con esta idea en mente, ya entrados los años 50, comienza a trabajar en el terreno del discurso (en sus inicios) y de la lexicografía, quien sin duda podemos considerar como el padre de la lingüística de corpus: John Sinclair (1933-2007). Fue él quien, de la mano de varios colaboradores, creó los primeros textos sobre la elaboración, el análisis y la explotación de corpus textuales. El más importante de sus trabajos es el COBUILD (Collins Birmingham University International Language Database) del cual se desprende el ya famoso *Collins COBUILD English Language Dictionary* de 1987, cuya última edición salió a la venta en 2006. A partir del proyecto COBUILD el profesor Sinclair se adentró en los estudios sobre la forma de procesar los textos que son tomados en cuenta para la creación de diccionarios y aportó a la lexicografía un método que ya se ha generalizado, a saber:

> …el uso de grandes recopilaciones textuales, debidamente planificadas, ordenadas y codificadas, como el instrumento ideal del lexicógrafo para detectar voces y acepciones nuevas, o incluso para identificar usos que la lexicografía tradicional no había recogido. Naturalmente, la utilización de los corpus fue posible porque había surgido otra herramienta que facilitaba su aprovechamiento: el ordenador y su capacidad para procesar la palabra electrónica. (Sánchez, 2008)

Así, podríamos decir que el nacimiento de la lingüística de corpus como tal se da en los años 60 con el advenimiento de las computadoras y su uso poco a poco generalizado[2].

---

[2] De hecho, actualmente es poco probable que el estudioso de la lengua pretenda crear un corpus de estudio sin la ayuda de la computadora y las herramientas de análisis lingüístico disponible, de alguna u otra manera.

Además, hay que contar la influencia que tuvo la incursión de las cada vez más grandes colecciones de textos que se utilizaron en todas las áreas de la lingüística teórica (discurso, semántica, gramática, fonética, etc.), y que se asienta en 1987 con la aparición del primer trabajo de gran envergadura desarrollado a partir de esta metodología de trabajo, el COBUILD.

En México, la lingüística de corpus aunque no ha resultado ser un tema de estudio acogido con gran entusiasmo por diversas razones. Podemos contar con trabajos importantes en esta área como el *Diccionario del Español de México* del cual hablaremos más adelante, aunque su impacto no ha sido como el de áreas más llamativas para el lingüista mexicano como la gramática (sincrónica y diacrónica), en particular la sintaxis; la fonética y, en menor grado aunque representativas, la semántica y la morfología. De tal manera que aun cuando todas ellas necesitan valerse de la lingüística de corpus para desarrollar sus investigaciones, el alumno promedio actual está más interesado en el área de la lingüística teórica que en la de herramientas y la metodología para llevar a cabo una investigación de esta naturaleza.

Existen, empero, interesantes investigaciones y desarrollos basados en corpus, como el DEM, el cual es coordinado por el Dr. Luis Fernando Lara del COLMEX-quien es considerado como uno de los pioneros en la rama de la lexicografía computacional por la metodología utilizada en el proyecto que iniciara en 1973 bajo la tutoría del Dr. Antonio Alatorre- y que ha desembocado en otros interesantes y reconocidos diccionarios como el *Diccionario Fundamental del Español de México* y el *Diccionario del Español Usual en México*. Otro diccionario importante basado en estudios de lengua con ayuda de la lingüística de corpus y de la computadora es el DIME (Diccionario Inicial del Español de México) de 2003, cuyo responsable fue el Dr. Raúl Ávila, también del COLMEX.

En particular en la UNAM la lingüística de corpus no fue trabajada como tal sino hasta la llegada del Dr. Gerardo Sierra en el año 1999 con la creación del Grupo de Ingeniería Lingüística cuyas áreas de investigación tienen que ver directamente con los corpus lingüísticos y la cual, de hecho, es una rama central. No queremos decir con esto que en México no se hayan llevado a cabo investigaciones basadas en corpus o que estos no hayan sido construidos bajo criterios estrictos de conformación. Lo que se quiere dar a

entender es que hasta 1999 no existía en el territorio mexicano alguien que se dedicara exclusivamente a dicha tarea. Todos los expertos que han trabajado desde hace muchos años en investigaciones *basadas en corpus* se apoyaban en pocos textos teóricos sobre esta rama y atendían más bien a su intuición de hablantes especializados del español de México.

A partir de esta fecha se comenzó a desarrollar la lingüística de corpus de manera más formal con la creación de dos corpus que ya se pueden consultar desde Internet: el *Corpus Lingüístico de Ingeniería*, **CLI** (Garduño, Sierra, & Medina, 2004), y el *Corpus Histórico del Español en México*, **CHEM** (Medina & Méndez, Arquitectura del Corpus Histórico del Español de México, 2006); además se comenzó a diseñar y construir la primera versión del *Corpus de las Sexualidades en México*, **CSMX** (Medina & Sierra, 2004).

Actualmente se trabaja en la ampliación y refinamiento del CSMX, de lo cual se hablará en el capítulo 3 de esta tesis, y en la construcción un corpus sobre adquisición de la lengua materna en niños de edad escolar primaria, en conjunto con el *Instituto de Investigaciones Filológicas* de la UNAM y la Universidad Autónoma Metropolitana-Xochimilco; además se está creando el *Corpus Científico del Español de México*, **COCIEM**, como parte de la colaboración interinstitucional con el Colegio de México.

2.2. Clasificación de corpus lingüísticos informatizados

La clasificación de los corpus lingüísticos ha sido una ardua tarea entre quienes intentan dar un criterio homogéneo que dé fe de todas las posibles distinciones que una persona puede hacer al momento de diseñar su material de trabajo. Para nuestro caso hemos tomado la clasificación hecha por (Sierra & Rosas, 2008) quienes afirman que la clasificación de un corpus lingüístico textual e informatizado tiene que atender a los criterios que mencionamos abajo.

2.2.1. Según el origen de los textos (modo)

Este es el punto de partida de la clasificación de los corpus. De acuerdo con estos autores hay dos fuentes principales: la oralidad y los textos, de tal manera que se llamarán corpus

orales y textuales, según sea el caso. El segundo de ellos tiene como característica ser todo aquel corpus que encontramos constituido por escritos que han sido informatizados. Por su parte, los corpus orales tienen una subdivisión que integra a los corpus orales sonoros y los corpus orales transcritos. Los de naturaleza sonora son todos aquellos que se conforman por grabaciones de cualquier tipo y los transcritos atañen a las transcripciones gráficas de grabaciones de lengua oral. En este sentido, para el español contamos con el corpus DIMEx100 que contiene grabaciones del español hablado en México y cuyo responsable directo es el Dr. Luis A. Pineda del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS**)** de la UNAM. (L.A. Pineda, 2009)

2.2.2. Según la espontaneidad

En este punto se debe de tener en cuenta la fuente de la información que contendrá el corpus de acuerdo con el nivel de premeditación del informante. Así, un ejemplo de corpus oral *premeditado* será aquel que se realice siguiendo un guión o una lectura dirigida, mientras que las grabaciones de una charla sin tener de por medio algún otro recurso que oriente dicha plática será clasificado como *no premeditado*. De igual manera, en los corpus textuales obtendremos un corpus premeditado cuando se obtengan documentos escritos de medios impresos o publicados en Internet sin modificación alguna de terceros, mientras que las conversaciones de un chat o los comentarios de un foro de discusión, por ejemplo, serán parte de un corpus textual no premeditado.

2.2.3. Según su codificación o anotación

La anotación o codificación es un proceso –generalmente electrónico- que consiste en agregar elementos (marcas) de diversa naturaleza al texto del corpus que se desea trabajar. Así, si se desea analizar las partes de la oración de un corpus se tendrá que utilizar una anotación del tipo POST (Part of Speech Tagging), hay anotaciones de tipo semántico (Fellbaum, 1998), suprasegmental (Garrido Almiñana, 1996) y morfosintáctica (Medina & Méndez, 2006), entre otras. El corpus anotado será el que contenga algún tipo de estas codificaciones y el corpus simple será aquel que se conserve en texto plano o en

formato .txt (ASCII), es decir, que no tenga algún tipo de modificación que aporte información extra.

2.2.4. Según la especificidad de los textos

De acuerdo con este punto, los textos que conforman un corpus pueden provenir de fuentes que mantengan un perfil de especialización (corpus especializados) y también hay los que recogen todo tipo de textos sin importar el nivel social, cultural o académico al que pertenecen (corpus general). Dentro de la clasificación especializada podemos hallar aquellos de textos informativos y de textos literarios.

2.2.5. Según la autoría

En este caso, si todos los textos de la colección corresponden a un solo género literario se llamará corpus genérico, en cambio, si los textos proceden de un solo autor se denominará corpus canónico. Ahora, si no corresponden a ninguna de estas dos subclasificaciones se llamará corpus de autoría variada.

2.2.6. Según el tiempo

Para poder clasificar un corpus dentro de un periodo determinado se usarán los términos diacrónico y sincrónico. El corpus diacrónico será aquel que tome textos de varios lapsos y los compare o confronte; el corpus sincrónico es el que toma muestras de textos de un determinado periodo. Como bien se ha señalado en el texto antes citado, no se deben confundir estos términos ya que pueden existir corpus sincrónicos de algún tiempo pasado específico y corpus diacrónicos de, por ejemplo, el siglo XX.

Dentro de los sincrónicos encontramos una subdivisión que atiende a corpus contemporáneos y corpus periódicos. Los contemporáneos se componen de textos actuales, mientras que los periódicos comprenden textos de un tiempo pasado determinado.

Por su parte, los diacrónicos se llamarán históricos o cronológicos si atienden a una comparación entre diversos periodos o si están ordenados de forma cronológica, respectivamente.

2.2.7. Según el propósito

Actualmente hay pocos corpus que se hacen por el simple hecho de tener un conjunto de textos ordenados de alguna manera y sirven para un sinnúmero de análisis y motivos distintos (que dentro de esta clasificación son llamados multipropósito). Por otro lado, tenemos aquellos que están destinados a ser utilizados con un fin específico y que son la mayoría de los ahora diseñados y creados. Estos últimos, por su naturaleza de preelaboración son llamados *de propósito específico*.

2.2.8. Según el lenguaje

El punto al que se refiere es sencillo. Mientras los corpus monolingües son los que contienen texto en una sola lengua, los multilingües tendrán en más de un idioma o dialecto los textos que lo conforman.

Ahora, el corpus monolingüe puede estar constituido de textos originales o de traducciones de otros idiomas. Cuando se da el caso de que el corpus contenga los textos en su lengua original aunado a una de sus traducciones se llamará corpus paralelo.

2.2.9. Según la cantidad de texto

Para esta distinción se toman en cuenta cinco subdivisiones: corpus grande, corpus de referencia, corpus léxico, corpus monitor y corpus pequeño. El primero de ellos será aquel que contenga una cantidad amplia de textos en comparación con sus colaterales. Esto puede ser relativo ya que hace apenas unos años se consideraba grande a un corpus que guardara en sí un millón de palabras mientras que hoy en día un corpus grande tiene en su haber varios millones de ellas. Así, podríamos decir que lo grande de un corpus se modifica con el tiempo y este criterio atiende a las capacidades de almacenamiento de los recursos que sostengan a dicho corpus.

El corpus de referencia es el que toma sólo fragmentos de los documentos que se desean estudiar y no su integridad. Este tipo de corpus sirve para dar fe de la lengua de manera exhaustiva, ya que intenta recopilar todos los textos posibles. Aquí es muy importante tener en cuenta los rasgos de equilibrio y representatividad –que explicaremos

en el siguiente apartado-, ya que se trata de una colección mayor pero muy bien articulada.

El corpus léxico trata de recoger la mayor información posible acerca de las variaciones léxicas de una lengua, para lo cual recoge fragmentos de longitud definida de la lengua que se desea estudiar.

Por su parte, el corpus monitor trata de recoger material actual constantemente, es decir, incluye continuamente nuevos fragmentos de texto de una lengua y deshecha los más antiguos. De esta forma se logra tener un corpus sincrónico que se mueve a la par del tiempo. Este rasgo, de temporalidad, es su característica principal.

Finalmente, el corpus pequeño es aquel que no intenta cubrir demasiadas necesidades y generalmente está hecho para fines específicos. Así, se puede considerar como un corpus pequeño, por ejemplo, una edición de la obra de un autor que se va a utilizar para el desarrollo de una tesis.

2.2.10. Según la distribución del tipo textual

Aquí es de vital importancia contar con el porcentaje de los diversos tipos de textos que son incluidos en el corpus. En la etapa de preparación del corpus se debe definir este punto ya que será el criterio que nos dirá si el corpus es equilibrado o no de acuerdo con su distribución. Otro de los puntos importantes dentro de la distribución es el que nos dice que podemos definir niveles, es decir las secciones en las cuales estará dividido el corpus. Ahora bien, existen también corpus piramidales que enlazan los criterios de variedad, y nivel con la distribución. Un corpus piramidal será aquel que en alguno de sus niveles contenga una mayor cantidad de textos y, por lo tanto, alcance mayor variedad; en el siguiente nivel puede contener una menor cantidad de textos pero más variados y así sucesivamente.

2.2.11. Según la accesibilidad

Lo importante en este punto es saber si el corpus estará disponible para el público en general o si será de uso restringido. El corpus de dominio publico puede ser comercial o

no comercial, esto quiere decir que se cobrará una cuota cada vez que se necesite consultar o una membresía que abarque periodos de distinta duración.

La restricción se debe, en muchos casos, a que los corpus son generalmente trabajos de investigación que han hecho grupos o instituciones y se debe de hacer un registro de usuarios para que no se haga mal uso de dicho material. La accesibilidad se debe también al soporte en que estará resguardado el material; este puede ser un sitio de Internet, un CD, un libro, etc.

2.2.12. Según la documentación

Este punto tiene que ver directamente con el aspecto legal de nuestro corpus. Si bien no se está acostumbrado a pedir todos los copyright de los documentos que son incluidos en los corpus -porque muchas veces no se conoce al encargado legal de estos o porque simplemente son textos bajados de Internet- sí es necesario hacer como mínimo una documentación sobre las fuentes de las cuales los textos utilizados para conformar el corpus fueron extraídos. Así, un corpus documentado será aquel que cuente con esta información y será un corpus no documentado aquel que omita completamente la información que hay detrás del texto extraído.

2.3. Criterios para la conformación de un corpus lingüístico

El diseño y la creación de un corpus lingüístico conlleva muchas más tareas de las que se podría imaginar y es, de hecho, la parte que más tiempo lleva en una investigación lingüística si es que esta contempla contar con material original desde su inicio. Hacer un corpus con varias personas a cargo tiene el fin de que las metas a las que va destinado puedan ser cumplidas de manera satisfactoria, por eso la tendencia indica que actualmente los corpus son diseñados por un grupo de personas y no por individuos.

He aquí el punto inicial que se debe tomar en cuenta para establecer los criterios que se seguirán en la conformación de un corpus lingüístico: siempre la colección de textos tendrá una inclinación marcada por el tipo de tarea a la que va destinada su creación. Así, un corpus que será utilizado con fines médicos dará preponderancia a los teóricos más actuales e importantes a la vez, mientras que un corpus con fines de estudio

lingüístico se inclinará, por ejemplo, hacia la representación del habla culta o del habla popular de una zona específica. Queremos decir que lo primero que hay que tener en mente al diseñar un corpus es saber hacia qué tarea va dirigido en primer lugar; si este es después consultado con otros fines será un motivo extra de satisfacción pero no un rasgo definitorio.

Ahora, como menciona Sierra (2006) un corpus lingüístico está destinado a ser siempre una muestra de la lengua y no un reflejo de la totalidad de ella, ya que cualquier corpus, por muy grande y ambicioso que sea, no puede abarcar la inmensa variedad de registros, cambios, fenómenos o grupos sociales existentes en una comunidad. Sin embargo, este es uno de los puntos clave en la conformación del corpus y de aquí parte la clasificación y la metodología propuesta. Se intenta, con base en este premisa, crear un corpus que sea representativo pero que al mismo tiempo tenga una amplia cobertura y que no se incline hacia algún parámetro, es decir que sea balanceado. A continuación definimos cada uno de estos rasgos.

2.3.1. Representatividad

La representatividad es el criterio principal que se debe de seguir para que un corpus tenga un mayor éxito como recurso lingüístico. Con esto queremos explicar que si bien intentar representar toda la lengua –española, alemana, inglesa, etc.- es una tarea poco menos que imposible, sí es posible hacer un inventario más o menos amplio que abarque el mayor número de rasgos posibles de la lengua en cuestión. Para esto, es necesario tener en mente los diversos puntos que se deben acatar con el fin de que esta fase no se haga un cuento de nunca acabar.

En primer lugar tenemos que la representatividad en un corpus puede verse como una convergencia de textos provenientes de diversas localidades geográficas dentro de un mismo territorio nacional[3]. Así, en un corpus del habla de México, si deseamos que

[3] Aunque también puede atender a un criterio distinto. Por ejemplo, la variedad puede verse como un conjunto abarcable dentro de un grupo denominado lingüística. Para que un corpus de lingüística sea variado deberá tener el mayor número de textos de las diversas disciplinas que la conforman: semántica, sintaxis, pragmática, fonética, etc.

[4] Si tomamos en cuenta que podemos considerar como hablante de español a un no nativo que, por muchos años que lleve viviendo en México, si no es al menos de tercera generación, es posible que aun cuente con

representativo tendríamos que tomar en cuenta las muestras de cada uno de los estados y no sólo de los más importantes económicamente o más densamente poblados, por ejemplo. Este punto debe ser cuidadosamente planteado ya que puede llevarnos muchas veces a confusiones que acarreen información falsa[4] y, por tanto, una representación menos fidedigna.

En segundo lugar tenemos la fuente primaria (en sentido amplio) donde se extraen los textos: el informante. Generalmente esta distinción se hace siguiendo el precepto de "abarcar más con menos", es decir, si se cuenta con la información personal de los informantes que serán finalmente los emisores del texto o audio capturado para la elaboración del corpus, se podrá hacer un registro y un estudio más detallado en algunos aspectos. Así es como surgen por ejemplo las denominaciones de "habla culta" o "popular", pues se atiende a la escolaridad o nivel de conocimiento del informante; la de "edad escolar", que atenderá justamente a niveles marcados por el sistema educativo de cada país; la de "altiplano central", que prepondera la ubicación de los informantes -generalmente debido a la congregación de un mayor registro lingüístico en ciertas áreas-; "especializado", por campo de trabajo, etcétera.

En un tercer plano encontramos la cuestión inconfundible de *tópico*. Sierra nos dice que esta distinción será siempre subjetiva pero necesaria y, como podemos constatar y como lo hemos mencionado al principio de este apartado, en la construcción de un corpus lingüístico siempre será imprescindible la ayuda de al menos un experto en el tema en torno al cual girará dicho corpus, pues este especialista será quien con mayor precisión conozca la tipología del área y las fuentes más importantes para la extracción de documentos.

Otro punto importantísimo que se debe tomar en cuenta es el tipo de texto que se utilizará. Un texto oral y un texto escrito estarán casi siempre destinados a estudios muy

---

rasgos y vicios de su lengua materna. Incluso si no hay acento por parte de dicho hablante es muy probable que, sintácticamente o en su inventario léxico, encontremos dichas anomalías o costumbres.

distintos y, por tanto, los rasgos de cada uno deben estar bien documentados y diferenciados. El tipo de texto nos servirá para saber, por ejemplo, si es habla espontánea, si hay una sintaxis específica en un discurso, las figuras y recursos retóricos utilizados por ciertas personas, etcétera.

La fuente del texto es un dato que casi nadie toma en cuenta y que muchos solicitan constantemente. Colocar en algún lugar del corpus una lista de fuentes explícitas de donde fueron extraídos los textos le dará al corpus un mayor índice de credibilidad y, además, se podrá contar con relaciones entre autores, trabajos y consultante. Si atendemos a la organización del texto encontraremos que su soporte puede ser un libro, una página de Internet, un manuscrito, etc. Si por otra parte, atendemos a la forma en que fue obtenido un texto oral, podríamos nombrar a los programas de radio, las entrevistas o la televisión.

Quizá el punto que nunca desaparecerá para determinar la representatividad de un corpus sea el que tiene que ver con la ubicación de este en el tiempo. Ya sea sincrónico o diacrónico, el corpus lingüístico debe tener una referencia temporal bien definida con el fin de que los estudios hechos con base en él sean confiables y cumplan con su objetivo, es decir, que haya coherencia entre el estudio que se va a realizar (análisis sintáctico, léxico, semántico, etc.) y el objeto de estudio (la lengua documentada en el corpus consultado).

2.3.2. Variedad

La variedad es un criterio de alta importancia para los corpus porque de ella depende que el conjunto de textos refleje con claridad el universo de rasgos distintivos de una lengua en un territorio y tiempo determinados. Nos referimos a variedad cuando perseguimos la idea de que hay que documentar cada una de las variantes de la lengua que se esté estudiando para llegar a la meta de representatividad antes descrita.

En este punto cabe aclarar que si bien cada uno de los criterios para la conformación de un corpus tiene su particularidad y razón de ser, todos ellos están ligados íntimamente pues dependen del éxito del otro para consolidarse apropiadamente. Así, la variedad puede ser realmente esto siempre y cuando cumpla con el objetivo de

tratar de documentar las variantes de la lengua del corpus, pero está limitada a su vez por la noción de equilibrio –que se verá en el siguiente apartado- y es presionada hacia la expansión por el criterio de representatividad, el cual le exige cubrir cada vez más variantes o territorio lingüístico con el fin de cumplir su tarea "representativa".

2.3.3. Equilibrio

Este es el criterio que tiene que ver con los límites de lo que se toma en cuenta para un corpus y lo que no. Unido estrechamente con el tamaño del corpus, el equilibrio intenta hacer de la colección de textos una estructura homogénea que no se vea afectada o influenciada por alguno de sus subtemas o tópicos. Equilibrio significa poder contar con una muestra representativa y variada de alguna lengua siempre y cuando la distribución de los textos sea -en el mejor de los casos- equitativa tipológicamente o por las divisiones hechas anteriormente por el compilador y sus colegas.

3. Integración del Corpus de las sexualidades en México (CSMX)

3.1. Planteamiento del corpus de trabajo

Para poder crear nuestro corpus decidimos atender a todos los criterios antes mencionado de una manera sistemática para que la extracción de términos fuera exitosa. Nuestro corpus es de modo textual, es decir, contiene escritos y no grabaciones o imágenes. En cuanto a la espontaneidad es premeditado y no premeditado, debido a que los textos extraídos fueron guardados sin modificación alguna y fueron concebidos para ser publicados, pero también tenemos algunas opiniones de foros de discusión e incluso diálogos escritos entre debatientes en algunas páginas; con esto le damos riqueza léxica al corpus.

Para el punto de codificación fuimos un poco más minuciosos ya que aunque el corpus en un principio se puede decir que estaba anotado, lo limpiamos y lo codificamos de manera distinta para efectos de nuestra investigación. (ver puntos 3.2. y 3.3.). Asimismo guardamos ambas versiones ya que el corpus no anotado nos sirvió para la extracción de términos y el corpus anotado sirvió para alimentar la base de datos del CSMX que se puede consultar en Internet. Por tanto, tenemos un corpus codificado con una versión simple.

En cuanto a la especificidad, se trata de un corpus general, ya que contiene tanto textos especializados como textos de lengua coloquial. La decisión de crearlo bajo ambos criterios (especializado y no especializado), aunque tratamos de extraer una terminología –que a grandes rasgos significaría contar sólo con textos especializados-, fue que para poder determinar con un mayor índice de acierto la especialidad de los textos y la realidad lingüística de México, tendríamos que tener varios niveles que nos indicaran esta estratificación. Además, como ya es por todos conocido, el fenómeno de uso de términos en la lengua coloquial y la conversión de palabras simples en términos es muy común y no podría determinarse en qué momento un término deja de pertenecer a esta categoría; de la misma manera es poco probable determinar el momento exacto en que una palabra común se convierte en término. Así, con la conjunción de ambos tipos de textos buscamos la profundización en la búsqueda y la extracción de la terminología básica.

El tema central del corpus es uno de los más controvertidos y abordados por la sociedad en general. Atendiendo a esta premisa, no podríamos haber creado un corpus canónico debido a que hubiéramos caído en un grave error. Las fuentes y los autores de los textos que conforman el corpus son variados, por lo que podemos encontrar más de un registro y más de un punto de vista en él.

De acuerdo con el tiempo, nuestro corpus es de naturaleza sincrónica ya que abarca el periodo actual. Tomamos como "actual" desde el inicio de esta nueva etapa del CSMX: septiembre de 2007 hasta la fecha: agosto de 2010. Incluso debemos de mencionar que este proyecto está pensado para, al menos, ocho años más, durante los cuales iremos integrando nuevos textos. Aun así nuestro corpus puede conservar la categoría de sincrónico ya que se trata de documentos pertenecientes a los últimos diez y hasta 15 años.[5]

Por supuesto, y como ya mencionamos antes, los corpus están siendo diseñados y creados principalmente con fines específicos. Así, el nuestro no escapa a dicha aseveración ya que fue concebido como parte de un proyecto y ha sido desarrollado pensando en ser explotado para fines lexicográficos y terminológicos. Se desea, además, que pueda ser usado en ámbitos fuera de la lingüística –medicina, psicología, sociología, etc.- lo cual no le quita su carácter específico sino que amplía su impacto.

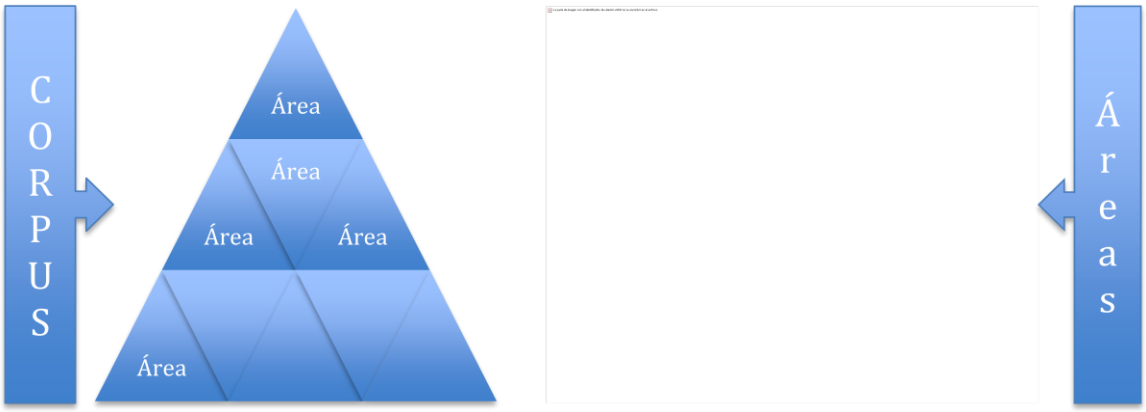
Como se menciona en el titulo de la presente tesis, el corpus fue diseñado para estudiar y extraer la terminología del español de México, por tanto es monolingüe e incluso podríamos decir que monodialectal ya que sólo abarca la variante mexicana del español.

Por la cantidad de textos extraídos hemos acordado en llamarlo corpus grande. Pero como bien ya hemos dicho, el tamaño puede ser un tema que se preste a discusión debido a que apenas cuenta con medio millón de palabras y un corpus grande, actualmente, podría contener arriba de los diez millones de ellas, por tanto debe la

[5] Para mayor referencia se puede consultar el Apéndice A de esta tesis con el fin de identificar las fechas de recuperación y de publicación de los textos utilizados para crear el corpus.

diferenciación en esta categoría por no ser aquel léxico, de referencia o monitor pues su destino y uso son distintos.

La distribución de nuestro corpus fue piramidal, es decir, que de un total de 160 archivos los hemos dividido en las ocho subáreas de la sexualidad y éstas a su vez en cinco niveles que contienen cuatro archivos cada uno. Un esquema de esta distribución es el siguiente:



Para la accesibilidad decidimos dejarlo totalmente libre de cualquier pago por consulta o uso y, por supuesto, está disponible para cualquier persona que desee utilizarlo con fines de investigación. Este punto fue resuelto así ya que, en primer lugar, no contamos con algún registro que nos de derecho a lucrar con este material porque no somos dueños de los textos en ningún caso, y en segundo lugar, quizá el más importante, porque se trata de un producto derivado de una investigación científica que está al servicio de todo aquel que lo necesite. Además, con esto respetamos una de las ideas principales que persigue este proyecto: la información detallada y confiable sobre la sexualidad en México en pro de la educación y la difusión de la ciencia.

Finalmente tenemos que nuestro corpus es un pieza documentada de investigación. El informe detallado sobre cada uno de los archivos que hemos integrado se encuentra en el Apéndice A de esta tesis. Asimismo, se pretende integrar como un apartado en la segunda versión del CSMX en línea para que cualquier persona tenga acceso a esta base de datos.

3.2. Criterios para la conformación del CSMX

El primer criterio, que tiene que ver con una división del gran tema de la sexualidad en subáreas, fue hecho atendiendo a los estudios de los expertos del Instituto Kinsey de Sexualidad con el fin de que nuestro corpus tuviese un alcance óptimo no sólo en una parte de lo que constituye esta área biológico-social, sino que además se apegará a la realidad lingüística completa del área de especialidad, pero sin olvidar el léxico coloquial. En otras palabras, para el rasgo de variedad, que está íntimamente ligado a la representatividad, se tuvo que caer en la cuenta de que la primera condiciona a la segunda cuando la diversidad de textos que se buscan se ligan por un lado al propósito del estudio al que va dirigido el corpus y por otro a los estratos que fueron tomados en cuenta para la extracción de dichos textos. Así, esta variedad abarca dentro de sí criterios diasistemáticos como el diatópico (de la lengua en la que se trabaja con respecto a su variación geográfica), que en nuestro caso atiende a toda la región que conforma México; y el diatécnico (la variación léxica de acuerdo con hablantes de la misma área de especialidad o sociolecto) puesto que esta tesis trabaja principalmente con una orientación terminológica.

El segundo criterio, la representatividad, se resolvió al definir que la terminología básica estaría orientada a describir el español mexicano de acuerdo con una división en niveles o marcación distrática (diferenciación por grupo social). Lo que desembocó en una clasificación por el origen de los documentos: academia, foros, asociaciones, etc., cada uno de los cuales delimita su contenido por el tipo de hablantes que lo conforman.

Finalmente, el tercer criterio tiene que ver con el equilibrio y se resolvió, atendiendo a Sierra, con la construcción de un corpus más apegado a los puntos de representatividad y equilibrio que al de tamaño. Lo que quiere decir que todos nuestros textos, aunque no muy extensos, intentan cubrir el mayor número de registros posibles del léxico que se utiliza actualmente en México en la Web. Con estos puntos en mente se procuró diseñar una estructura de corpus estable que finalmente resolvimos de manera piramidal. Esta estructura comprendió la variedad cuando por cada subárea se armó una sección del corpus y a la representatividad cuando por esas mismas subáreas se otorgaron niveles y por cada nivel un número de archivos. Para que nuestro corpus estuviera equilibrado fue necesario establecer un límite en cuanto al tamaño de los archivos se

refiere con lo cual se evitó que textos muy extensos influyeran demasiado en la extracción de los términos que se utilizaron para la conformación de la terminología básica.

3.2.1. La variedad del corpus: áreas temáticas

Como ya mencionamos, la variedad estuvo constituida por los parámetros que dicta en Instituto Kinsey. De acuerdo con esta institución el área de la sexualidad está dividida en ocho apartados y cada uno de ellos representa una visión distinta y necesaria. Los apartados son los siguientes:

Fundamentos biológicos de la sexualidad. Es el área donde se insertan los documentos relacionados con la anatomía humana, el cuerpo visto como objeto de estudio, el control de la natalidad y los documentos relacionados con la medicina general que abordan temas de sexualidad.

La respuesta y la expresión sexual. Tiene que ver con asuntos metabiológicos encontrados en la mayoría de los documentos (razones sociales, cuestiones psicológicas, etc.). Incluye estudios sobre el orgasmo, las fantasías sexuales y los afrodisíacos, entre otros.

Comportamiento sexual. Se trata de documentos relacionados con la masturbación, la actividad sexual no coital, la estimulación oral y anal, el coito, la sexualidad postural y el aumento o disminución del placer en las relaciones sexuales.

La identidad sexual. Textos referentes a la identidad del sujeto, los roles y la orientación sexual, los fundamentos biológicos de la identidad sexual, las fases del desarrollo de la identidad sexual, los roles y estereotipos masculinos y femeninos, las anomalías en el desarrollo de la identificación sexual y temas referentes a género.

Las enfermedades de transmisión sexual (ETS). Documentos dedicados a la amplia gama de enfermedades actuales transmisibles vía sexual –desde su sintomatología hasta los análisis profundos publicados recientemente-, el control de estas enfermedades y los mecanismos de prevención.

La sexualidad variante. Aquí se agrupan los documentos referentes a los trastornos y las diferencias psicosociales en el área de la sexualidad: fetichismo, travestismo, masoquismo y sadomasoquismo, actividades poco documentadas (coprofilia, urofilia, saliromania, zoofilia, etc.); el tratamiento de las llamadas parafilias (conductuales, técnicas bioquímicas, etc.), además de documentos que definen y clasifican estos comportamientos.

Atracción sexual y relaciones en pareja. Aquí entran documentos que versan sobre el origen y el desarrollo de las relaciones de pareja, las habilidades sociales en la conquista de la pareja, temas sobre fidelidad e infidelidad, conceptos biológicos y culturales de las relaciones de pareja, función de las relaciones en pareja, las terapias de pareja, etc. Así como temas sobre el abuso u ofensa sexual.

La educación sexual, cultura y sexualidad social. Documentos sobre los procesos de la educación sexual (escuela, familia, sociedad, etc.), la educación sexual por edades, la sexualidad y sus limitaciones, la relación sexual en el tiempo y en la cultura occidental, las habilidades sociales en el comportamiento sexual y la sexualidad en diferentes culturas.

3.2.2. La representatividad del corpus: niveles léxicos

La representatividad se busca en un corpus con la finalidad de incorporar el mayor número de posibilidades que el hablante de una lengua tiene acerca del léxico que domina. En este caso, los niveles que a continuación presentamos y que utilizamos para dividir nuestro corpus intentan cubrir el mayor número de registros posibles de los hablantes del español de México.

Nivel 1: Documentos de Google académico. El buscador Google tiene entre sus opciones de búsqueda un apartado denominado *Google Académico* cuya finalidad es proporcionar a los interesados sólo textos de especialistas en los temas de las palabras introducidas en el motor de búsqueda. Este nivel es el más alto debido a que se trata del registro de los especialistas en el tema de la sexualidad y son textos reconocidos con un mayor número de referencias de acuerdo con este buscador. Dentro de ellos se encuentran publicaciones

variadas que van desde artículos arbitrados hasta libros completos que versan ampliamente sobre el tema que nos atañe.

Nivel 2: Documentos de asociaciones sobre el tema. En este rubro se integran los textos extraídos de páginas de asociaciones relacionadas con los temas de sexualidad, tales asociaciones pueden ser públicas o privadas. Este nivel es el segundo ya que de él se desprenden textos que integramos con el conocimiento de que la fuente de la que provienen tiene reconocimiento en el campo de la sexualidad en México; se hallaron textos de instituciones educativas, de salud y de servicios que dependen de programas de desarrollo social.

Nivel 3: Artículos PDF. Son todos aquellos documentos que están protegidos, es decir, guardados en versión *PDF* para que no sean modificados. Se trata de textos de divulgación o investigación más seria que sin embargo no llegan a ser académicos por no estar respaldados por asociaciones, universidades o centros de investigación de prestigio. Su ubicación dentro de nuestra clasificación en niveles atiende a que muchos de ellos no están muy referenciados pero representan una visión respaldada por la experiencia de los autores en el área de la sexualidad como comunicólogos, difusores de la ciencia, etc.

Nivel 4: Html,Word. Este nivel contiene los documentos que sólo están disponibles en formato Word o en formato de página Web, es decir, en codificación HTML. Son textos de Internet que sólo fueron copiados y pegados en hojas de texto. Se trata de textos que encontramos en páginas Web dedicadas a la sexualidad que no cuentan con algún respaldo institucional pero que son muy visitadas y consultadas por una gran cantidad de usuarios de acuerdo con su posicionamiento en las listas de los buscadores utilizados por nosotros (Google, AlltheWeb, Lycos, Yahoo!, etc.)

Nivel 5: Foros. Finalmente, a esta sección corresponden los archivos de los foros de discusión reconocidos o completamente libres. Se cuidó que este apartado no coincidiera con los foros de preguntas y respuestas que contienen las páginas de asociaciones. Este nivel está constituido por textos que contienen en gran medida lenguaje coloquial y de uso cada vez más generalizado pero que no está documentado en los otros niveles. En este punto es donde podríamos ver un acercamiento a la oralidad debido a que si bien se

trata de texto, la forma en que está utilizado podría asemejarse a un acto de habla; así podemos ver que hay una línea difusa entre oralidad y escritura en nuestros tiempos.[6]

3.2.3. El equilibrio del corpus: tamaño de los documentos

El último de los criterios que se adoptó fue el equilibrio. El equilibrio tiene que ver íntimamente con la fuente primaria de los documentos (representatividad) y con la subárea a la que pertenece (variedad). Así, este criterio es el que menos características tiene al componerse de un solo parámetro: el tamaño. Cada uno de los archivos utilizados debió tener un peso máximo de 5Mb no importando si era un artículo, libro o texto de un foro de discusión. Además, como ya se mencionó, el número de archivos por nivel fue el mismo, así como el número de archivos por área. Se cuidó que el tamaño de los archivos correspondiera con su codificación, es decir, que este límite fue utilizado con documentos ya limpios (ver apartado 3.3.) debido a que la codificación HTML o similar aumenta considerablemente el tamaño de un archivo.

3.3. Fase de limpiado

Para que un corpus pueda ser procesado y explotado con fines lingüísticos, siempre es necesario que esté anotado de algún modo particular, con algún lenguaje específico o incluso debe estar totalmente limpio, es decir, en un formato de texto plano que para fines de esta investigación hemos convenido en llamar *formato.txt* –ya que es la extensión por omisión de los archivos en texto plano de Windows-.

Pues bien, ya sea que se necesite en texto plano o anotado, un corpus debe pasar por una fase de limpiado, ya que todo texto que ha sido subido a la Web necesariamente ha pasado por un proceso de anotación[7] con el fin de visualizar ciertos elementos,

[6] Es interesante hacer notar en este punto que lo mismo sucede con las conversaciones que son recuperadas de los salones de chat o las pláticas guardadas por los mensajeros instantáneos como *live Messenger, Yahoo, Googlechat, Adium,* etc. Una investigación interesante sobre este tema la podemos consultar en el trabajo de Eres Fernández y Almeida Seeman, Eres Fernández, I., & Almeida Seemann, P. (2009). "Un estudio sobre los cambios lingüísticos del español escrito en las charlas informales por Internet" en *Trabalhos em Linguística Aplicada , 48* (1).

[7] Para mayor referencia véase el artículo de Kilgarrif & Grefenstete (2003) "Web as a Corpus" en *Computational Linguistics*, Volume V.

delimitar otros, darle formato al texto, crear vínculos entre palabras, enlazar ciertas imágenes con explicaciones en texto e infinidad de características y funciones que la anotación nos da y que no vemos directamente cuando abrimos una página Web (que es lo que se llama en este ámbito el *código fuente*).

Limpiar significa eliminar todas las etiquetas que agrega la codificación HTML en que están las páginas Web, con el fin de integrar las que nosotros hemos definido y para crear la base de datos y el soporte de consulta. El método consiste en bajar toda la información de la página consultada, extraer el documento en texto plano con etiquetas y eliminarlas de forma semiautomática con la ayuda de las herramientas que permite utilizar el procesador de textos elegido -que en nuestro caso se trató del blog de notas incluido en la paquetería de Windows-. Esta versión del documento, ya sin anotación, es la que utilizaremos para procesar el corpus.

4. Herramientas para el procesamiento del CSMX

Una vez que tuvimos el corpus completo y limpio pasamos a la fase de procesado para que la extracción terminológica concluyera con éxito. El procesamiento y explotación de un corpus tiene diversas perspectivas de acuerdo con la finalidad de la investigación para la cual fue creado. En cualquiera de los casos, es necesario que los documentos que conforman el corpus de trabajo estén etiquetados, así, una vez planeado, diseñado y recolectado el corpus, esta es la tarea imprescindible que hay que llevar a cabo. Cabe mencionar que para ciertos investigadores este paso es omitible o innecesario debido a la naturaleza de su trabajo, para nosotros no es así. Ya que trabajamos con unidades léxicas altamente especializadas y que necesitan ser analizadas como parte de un lenguaje específico y no como entidades aisladas –ya sea estadísticamente o desde el punto de vista semántico- debemos de poner en claro qué tipo de unidades son éstas. Más adelante veremos que un término no es sólo una palabra gráfica que tiene ciertas relaciones con otras por cuestiones de aparición junto a ellas o por algún tipo de nexo que las une, sino que forman parte de un solo constructo con significado único determinado por el contexto, en muchos casos.

4.1. Etiquetado

El etiquetado o anotación en lingüística de corpus es el proceso de asignar *tags* o etiquetas a cada uno de los elementos que lo conforman –palabras gráficas-, es decir, nombrarlas, ya que la computadora sólo lee las palabras como cadenas de caracteres y no como segmentos con cualidades semánticas. En una pocas palabras, se crea de forma codificada el metalenguaje que el hablante de una lengua adquiere de forma natural desde que nace: la gramática.

Basándonos en Leech (1997) podemos decir que un corpus anotado podrá tener las siguientes características de acuerdo con el tipo de estudio al que va dirigido. Es importante mencionar que habrá que cumplir algunos principios como los que se enumeran a continuación para que la anotación tenga una efectividad considerable:

- Debe ser posible remover la anotación de un corpus y convertirlo en corpus no anotado.

- Debe ser posible extraer las anotaciones de un corpus y ser guardadas de manera independiente.
- Las etiquetas deben estar basadas en documentación disponible para el usuario, que incluya el esquema de anotación; e incluso dar información sobre la confiabilidad y consistencia de la anotación seguida.
- Debe dejarse claro cómo y por quién fue realizada la anotación.
- Debe darse información sobre la confiabilidad y consistencia de la anotación seguida. El usuario final debe estar consciente de que la anotación de un corpus no es infalible, sino simplemente una herramienta poderosa.
- Los esquemas de anotación deben estar basados en principios ampliamente definidos, de preferencia en consenso y en teorías neutrales.
- No deben considerarse los esquemas de anotación como una estándar, ya que tienden a variar por razones prácticas.

El etiquetado es, en pocas palabras, la gramática detrás del texto que se está procesando, es lo que nos dice si una palabra es un sustantivo, un verbo, un adjetivo, etc.

Nuestras etiquetas marcan, entre otras cosas, los datos del autor: nombre, dependencia a la que pertenece; datos del texto: título, subtítulos, pies de página y dirección de Internet donde fue encontrado. Como podemos ver, dichas entidades de marcaje ya no sólo están indicando categorías gramaticales, sino que también marcan cualidades semánticas más allá de restricciones de tipo estructural o funcional; en ellas marcamos elementos que son importantes para nuestro trabajo y van desde la identificación de las funciones de una palabra dentro de cada documento hasta rasgos mínimos de cada categoría gramatical como género y número.

4.2. Herramienta Tok

La Herramienta Tok es un programa desarrollado en el Grupo de Ingeniería Lingüística para preprocesar el texto plano y codificarlo en XML. La ventaja que ofrece esta herramienta es que nos ahorra mucho tiempo en la fase de etiquetado. Su función

principal es separar las palabras y crear etiquetas vacías que sólo necesitan ser llenadas con las convenciones de marcaje que un proyecto requiera.

Para fines de esta investigación, la herramienta Tok hizo las veces de una guía de marcaje para el etiquetado manual con el fin de que, en la fase de creación del soporte de consulta, cada uno de los elementos que se determinaron clave tuviera una relación directa con la totalidad del corpus; es decir, una tarea mecánica para delimitar una palabra, un signo de puntuación o una formula -entre otros elementos que se pueden encontrar en un texto- fue hecha de manera automática. Particularmente, la herramienta Tok tomó el conjunto de textos limpios, separó cada palabra gráfica y le asignó una etiqueta vacía que contiene datos morfosintácticos básicos de tal manera que el etiquetador humano sólo se concentró en aplicar sus conocimientos lingüísticos para determinar los posibles errores y corregirlos sin necesidad de anotar toda la información manualmente aparte de asignar la etiqueta correcta para cada palabra a la par de especificar si un documento está bien formado.

4.3. XML (Extensible Markup Language)

XML es un lenguaje, esto es, un código, un conjunto acordado de signos para la comunicación. Pero es un lenguaje de marcado, lo que significa que sirve para modificar el significado de otros símbolos para dotarlos de mayor significado. Es extensible, de modo que el conjunto de símbolos no es fijo, sino que puede ampliarse para que pueda abarcar prácticamente cualquier ámbito en el que sea preciso identificar —marcar— cualquier tipo de información. A diferencia de HTML (HyperText Markup Language: "Lenguaje de Marcado de Hipertexto"), su antecesor, no tiene elementos predeterminados para marcar los componentes típicos de una página Web, sino que permite generar elementos propios y asignarles los nombres de nuestra preferencia. En realidad, tanto XML como HTML provienen de un lenguaje más antiguo, llamado SGML, o Lenguaje Generalizado de Marcado Estándar. SGML es, en realidad, un meta-lenguaje, o lenguaje para describir lenguajes. Se utilizó en los años 60-70 para crear sistemas de gestión documental, y descripción de datos, pero la especificación era tan enorme y había tantas sutilezas en su uso, que al final no ha tenido más utilización práctica que describir un par de lenguajes, HTML y XML.

En primer lugar se creó HTML, que era un conjunto de etiquetas para controlar la presentación de información en la entonces incipiente WWW. Posteriormente se creó el XML como subconjunto de SGML, de modo que XML también permite describir otros lenguajes, a través de ese modelo de datos, pero normalmente se utiliza para marcar, describir información que puede existir en muy diferentes formatos.

Como heredero de SGML que era, el lenguaje HTML estaba pensado fundamentalmente para "calificar" diferentes elementos de una página. Con ese archivo, el navegador creaba una presentación en pantalla de esos elementos, que podía variar enormemente de navegador a navegador. Cuando la Web pasó de simple herramienta de acceso a la información a un medio de comercio, HTML pasó de ser un lenguaje de marcado de contenido a ser un lenguaje de presentación. XML es ahora, además de un lenguaje de presentación, uno de descripción y ordenación.

Al ser un lenguaje de descripción, un meta-lenguaje, se está utilizando para muchas más cosas, especialmente aquellas que implican intercambio de información en un formato lo más autocontenido posible. De esta manera se pueden representar partituras musicales, ecuaciones matemáticas, inventarios de libros, estructuras jerárquicas, entre otras cosas, para las que HTML carece de elementos. XML ha desencadenado desde su aparición un fenómeno sin precedentes debido a su facilidad de uso y a la cómoda interacción con los usuarios. Se ha vuelto el lenguaje de marcado por excelencia ya que con él se pueden describir datos propios, moldearlos y controlarlos con un mínimo esfuerzo y con un conocimiento básico sobre cuestiones de programación y sintaxis. Es muy efectivo debido a que es prácticamente texto y porque además de ser reconocido por casi todos los navegadores, es posible modificar sus etiquetas a voluntad. Entre sus características se cuentan:

- Utiliza etiquetas, entre llaves angulares < >, para marcar, esto es, calificar, la información a la que rodea.

- Hay dos tipos de etiquetas, las de apertura, < >, y las de cierre </>.

- Las etiquetas no se entrecruzan, esto es, la última en abrirse siempre ha de ser la primera en cerrarse.

- El archivo XML es fácil de leer, no sólo para un ordenador, sino también para una persona.

La idea de utilizar este lenguaje de marcado es poder identificar cada una de las palabras de los textos extraídos de Internet para su correcto procesamiento en la interfaz de búsqueda que creamos. La ventaja de XML es que permite modificar los criterios de marcaje de tal manera que podemos utilizar una sola corriente lingüística o computacional, o mezclar varias para hacer más fino el análisis del texto.

Para crear un documento XML hay que considerar las siguientes reglas básicas:

1. El documento debe tener exactamente un elemento de nivel superior, y todos los demás elementos estarán anidados dentro de él.
2. Los elementos deben estar bien anidados, es decir, deben finalizar en el mismo elemento en el que inician.
3. Cada elemento debe tener un marcador de inicio y de fin, y el nombre de éstos debe coincidir.

Ahora bien, dentro de nuestra labor terminológica, en particular en el *Corpus de las Sexualidades en México* (CSMX), este lenguaje de marcado fue de gran ayuda ya que como se divide en ocho áreas (fundamentos, respuesta, comportamiento, identidad, ETS, sexualidad variante, atracción y educación) cada una de ellas contiene documentos organizados en cinco niveles diferentes y dentro de cada nivel se agrupan datos de cada artículo; una estructura jerárquica como esta hizo necesaria la utilización de un lenguaje que además de indicarnos las particularidades semánticas de cada palabra gráfica, también nos remitiera a la posición de esa palabra dentro de cualquier lugar del corpus. Un ejemplo de cómo organizamos estos datos se muestran en el cuadro de abajo y, como se puede ver, están correctamente anidados dentro de un solo elemento y cada marcador de inicio tiene uno de fin (donde *<etiqueta>* indica el marcador de inicio y *</etiqueta>* indica el final).

Con este conjunto de etiquetas fue posible generar relaciones entre las palabras ya con criterios lingüísticos y no únicamente probabilísticos o estadísticos. Así, cuando la búsqueda de términos comenzó pudimos podemos entender la naturaleza y el

comportamiento, por dar un ejemplo, de un término multipalabra o compuesto; con el marcaje en XML podremos saber de qué palabras está acompañado un término, cuál es la categoría gramatical de dichas palabras, en qué documento se encuentra, a qué área de la sexualidad pertenece, etc.

```
<Educacion>
<referencia id="csmx">
                    <titulo>Modelos de educacion sexual</titulo>
                    <autor>Gabriela Rodriguez Ramirez</autor>
                    <nivel>2</nivel>
                    <nombre_revista>Adolescencia</nombre_revista>
                    <num_publicacion>año 2, numero 10</num_publicacion>
                    <fecha_captura></fecha_captura>
                    <url>http://www.adolec.org.mx/litcien/boletin/bol10/Boletin10.pdf</url>
                    <CLAZ>educacion pdf 1</CLAZ>
</referencia>
</Educacion>
```

Ejemplo de etiquetado con XML

Para su correcta visualización en cualquier navegador de Internet, hay que revisar que un documento esté bien formado –que contenga todas las etiquetas de inicio y de fin correctamente anidadas- y que esté validado –que corresponda a una hoja de estilo- lo que significa que cada etiqueta cuenta con los valores correctos en su conformación. Si el documento está correctamente escrito (es decir, sin errores en los marcadores -con una sintaxis precisa-) aparecerá el documento en una ventana del navegador que utilicemos y, por lo tanto, los resultados pueden ser observados por los usuarios de la interfaz.

La interfaz es una de las varias aplicaciones que puede tener XML, el punto es que también está siendo utilizado en gran medida para los servicios Web y, al ser Internet un gran medio de difusión, decidimos poner en él nuestro servicio de consulta. Un servicio Web no es más que un programa que reside en un servidor remoto, al que se accede por medio de los mismos protocolos que utilizamos en los navegadores, pero que

recogen información que se les suministra y devuelven un conjunto de información relevante a partir de la información suministrada; así, el primer producto derivado de esta investigación es el generador de concordancias del CSMX.

4.4. Soporte de consulta

Tanto el corpus de las sexualidades como el vocabulario básico estarán disponibles en Internet. De hecho, como se mencionó arriba, el primero tiene ya una interfaz que permite al usuario buscar las palabras que necesita para su consulta y le despliega una lista de fragmentos textuales, es decir, la palabra y el contexto en que se inserta. Dicha interfaz se mejorará, haciéndola más productiva y atractiva al consultante. Por ejemplo, se agregarán los datos de la dirección de Internet donde se pueden encontrar los textos originales y los datos del autor del texto donde se extrae el fragmento que se muestra. Dicha interfaz puede ser consultada en la página: http://www.iling.unam.mx/csmx.

Interfaz actual del Corpus de las sexualidades en México

Una vez que se etiquetaron todos los textos fue posible pasarlos a la base de datos para actualizar la página del CSMX y, con los archivos limpios, procedimos a extraer la

lista de términos. Como podemos observar, esta fase del proyecto alimentó dos trabajos distintos y, por lo tanto, es el nodo central de la investigación. Con el corpus limpio y codificado y con las herramientas adecuadas podemos explotar nuestros recursos finalmente.

Con lo anterior queremos decir que tenemos dos versiones de un mismo corpus, la versión anotada que es la que utilizamos para el generador de concordancias que podemos consultar desde Internet y la versión limpia, que es la que utilizamos para la extracción terminológica con ayuda de las herramientas contenidas en el programa *Wordsmith.* Ambas versiones nos ayudan a identificar los candidatos a término y a obtener los términos mismos.

Esta parte de la tesis constituye la fase de final en cuanto a la preparación del corpus para una explotación terminótica eficiente y veloz. En este punto es donde se centra la agilización del proceso de elección de términos ya que bajo estos formatos (anotado y limpio) podemos aplicar muchas otras herramientas de análisis lingüístico que nos ayudan a reducir considerablemente el proceso terminográfico actual. Aquí podemos basar nuestras expectativas en cuanto a la rapidez con que un vocabulario básico o una terminología básica pueden ser creados. Mientras que en un proceso clásico la simple tarea de recopilación del corpus nos llevaría incluso años, con este procedimiento vimos que el proceso de recuperación de textos se reduce en cuanto al factor tiempo.

Ahora bien, con el uso de las herramientas que hemos descrito, la terminología básica puede ser diseñada y organizada en, a lo mucho, la mitad de tiempo que nos llevaría si creáramos fichas de trabajo y analizáramos manualmente cada texto del corpus incluso con el número de personas que trabajaron en este proyecto (alrededor de veinte).

Finalmente, podemos agregar que las herramientas creadas por el Grupo de Ingeniería Lingüística contribuyen a esta reducción de tiempo en cuanto a la anotación lingüística se refiere a la par de proporcionar un proceso más ágil y confiable para la obtención de unidades lingüísticas relevantes –palabras idénticas, palabras pertenecientes a ciertos paradigmas, lemas, etc.- en cualquier corpus, sean éstas o no términos ya que actúan con independencia del nivel de especialización de los textos de un corpus dado.

V. Extracción de la terminología básica de las sexualidades en México

Uno de los puntos que permiten establecer diferencias claras entre el lenguaje común y el especializado, como entre los distintos lenguajes especializados entre sí, es el uso de palabras que detallan o circunscriben un concepto dentro de un área. Estas palabras –conceptos- utilizadas en un contexto específico –área de especialidad- se llaman términos.

Es importante recalcar que en esta tesis intentamos extraer los términos que conforman nuestro vocabulario para, en etapas posteriores, encontrar las definiciones de ellos. Respecto a esto, la obtención de términos consta de tres fases que en conjunto son necesarias y suficientes para crear una lista de elementos básicos, o sea, un acercamiento a los términos más relevantes. Empleamos una metodología terminográfica apoyada en herramientas terminóticas, es decir, encontramos los términos a partir del corpus por métodos estadísticos automáticos con la ayuda de herramientas como etiquetadores, lematizadores, buscadores de palabras clave, generadores de listas de palabras, etc. e implementamos una evaluación y selección apoyada en análisis manuales.

Más adelante veremos cómo la extracción terminológica parte de una teoría terminológica y se aplica de acuerdo con el tipo y el área de trabajo de los encargados de cada proyecto de extracción.

5.1. Bases teóricas de la extracción terminológica

Para poder abordar correctamente en el tema de la extracción terminológica es vital hacer distinciones claras entre los conceptos que vamos a abordar: terminología, terminografía y terminótica.

Una distinción que hay que hacer cuando nos sumergimos en la extracción terminológica es que para poder separar la terminología y la terminografía de la lexicografía u otras ramas, es necesario decir que mientras la lexicografía define palabras o entradas, la terminología define términos, unidades especializadas. Esta rama, pues, necesita mucho de la semántica y está inevitablemente unida a la lexicografía, sobre todo por la metodología para la extracción de las unidades que se desean definir. Ahora bien, mientras los lexicógrafos elaboran diccionarios de lengua general –ya sean básicos, enciclopédicos, inversos, etc.- los terminólogos elaboran diccionarios especializados o de

lengua especializada –entre sus productos figuran vocabularios, glosarios, diccionarios terminológicos, etc.-.

5.1.1. Terminología

La terminología es el conjunto del vocabulario especializado de una disciplina o un ámbito de conocimiento. Es entendida como "la disciplina que estudia los términos, los conceptos y su relación"[8]. En esta investigación, el concepto de término lo enmarcamos dentro del contexto específico de la *Teoría Comunicativa de la Terminología* (TCT). Es importante aclarar que en la TCT "los términos son unidades léxicas activadas singularmente por sus condiciones pragmáticas de adecuación a un tipo de comunicación"[9]. Es así que en este trabajo consideraremos a un término como una unidad léxica especializada cuyo significado se encuentra relacionado con un área de conocimiento particular.

Esta es una palabra polisémica debido a que una parte de los teóricos denomina a dicha disciplina como meramente teórica pero también se usa para referirse a la práctica terminológica y, además, llaman así al producto de esta práctica, es decir diccionarios, glosarios, bases de datos etc., y sin embargo es llamada por otros terminografía. Para algunos, es una disciplina independiente, para otros se trata de una rama de la lingüística.

La terminología no es, en rigor, un campo de trabajo reciente: en el siglo XVIII el desarrollo de algunas ciencias y el advenimiento de la Revolución Industrial que afecto mucho lingüísticamente todas las áreas del conocimiento -pues las hizo más especializadas y estables- dieron origen a los trabajos de recopilación y ordenamiento terminológico de Lavoisier y Berthold para el caso de la química y los de Linnéo para el de la botánica, por ejemplo. Durante el siglo XIX, a raíz de la internacionalización progresiva de la ciencia, surge en el campo científico la demanda por establecer reglas de formación para los términos de sus disciplinas; en el siglo XX se suman a esta necesidad

[8] Cabré, T. (1999). *La terminología. Teoría, metodología y aplicaciones.* Barcelona: IULA.

[8] *Ibid*

las distintas ramas de la tecnología con mucha más fuerza, todas las cuales requieren orientaciones para denominar nuevos conceptos y, sobre todo, "armonizar", en el sentido de "regular y ordenar" las nuevas denominaciones con el fin de lograr una comunicación efectiva y eficiente.

En este contexto surgió la teoría general de la terminología, desarrollada por el ingeniero Eugenio Wüster, quien propuso la normalización conceptual y denominativa de los términos con el fin de hacer más efectiva y cristalina la comunicación entre los especialistas. Esto se da en 1931 cuando Wüster hace una tesis sobre términos electrotécnicos. Para 1938 elabora el diccionario *Machine Tool*.

Wüster sostiene que la terminología tiene varios padres: Schloman, quien elaboró un diccionario sistemático en seis lenguas; Ferdinand de Saussure, fundador de la lingüística estructuralista; E. Dressen, quien fue el impulsor de la ISA (International Standarization Asociation) -que fue la primera organización de normalización y actualmente se llama ISO (International Standarization Organization)- y finalmente J. Holmstrom, quien dijo que debía crearse un centro de estudio terminológico. Wüster así, es considerado el padre de la práctica terminológica, pero D. S. Lotte es el padre de la teoría terminológica según otros puntos de vista.

Para Rondeau (1984) la terminología es un fenómeno socioeconómico: el avance de las ciencias crea nuevos conceptos a los que hay que dar nombre. Algunos conceptos se conocen con diferentes nombres, esto crea problemas de comunicación y se soluciona con la univocidad de los conceptos, por eso la terminología es tan importante.

La terminología, como ya dijimos, surge en Austria, Unión Soviética y Checoslovaquia sobre los años 30. Se expande hacia el norte (Dinamarca, Bélgica) y el oeste (Francia, Canadá) durante los años 60. Más adelante llega al sur (España, Sudamérica, Portugal) y al este (China y Japón).

Tanto la escuela de Viena, fundada por Wüster, como la de Moscú y Praga son de tendencia lingüístico−terminológica (normalización, sistemas de conceptos). También están las escuelas traduccionalista y la normalizadora, de las cuales no hablaremos en este momento.

En nuestro país, la práctica terminológica y la teoría de la misma no han visto a muchos especialistas dedicados a desarrollarlas ya que esta rama entró hace relativamente

poco tiempo. Hacia 1975-80 la Dra. Ana María Cardero de la FES Acatlán desarrolla una tesis de licenciatura que tiene que ver con términos cinematográficos, con lo que da pie a la práctica propiamente dicha; mientras tanto, en el Colegio de México la Dra. María Pozzi comienza a trabajar con asuntos de teoría terminológica y normalización. Cabe mencionar que esta última es alumna de un grupo de terminólogos extranjeros que llegan a México justamente para integrar esta área al estudio lingüístico. Más tarde, en 1999, el Dr. Gerardo Sierra regresa a México después de hacer su doctorado en lingüística computacional y se relaciona con ambas doctoras, ya que él trabajo durante todo el posgrado en cuestiones de terminología pero apoyada por computadora. Actualmente ellos tres son los terminólogos consolidados y son quienes están formando a la nueva generación de esta área en nuestro país.

5.1.2.Terminografía

La terminografía es la práctica terminológica: recopilar, clasificar y representar términos. Los pasos preliminares y posteriores al trabajo terminográfico se denominan práctica terminológica: determinar a quién va dirigido, cómo se va a hacer, releer, etc. La terminografía es la parte del trabajo terminológico que se encarga de llevar a cabo los procesos de recopilación de información necesaria para el corpus de trabajo, la extracción y selección de términos a partir de un corpus y la conformación de los productos finales derivados de estas tareas.

En terminología se hace la distinción entre diversas entidades para poder entender el funcionamiento del lenguaje especializado ya que, de acuerdo con su naturaleza, los términos pueden ser agrupados en distintos campos.

Los términos simples están constituidos por un solo morfema y se forman por derivación o composición. Los términos complejos están constituidos por dos o más morfemas y forman un sintagma terminológico o, como lo podremos encontrar en este trabajo, términos multipalabra, es decir, una expresión con un sentido único.

Las abreviaturas son representaciones escritas resumidas de un término simple o complejo, tras suprimirse una o varias letras, sílabas o palabras del término original.

Los acrónimos son vocablos compuestos mediante el abreviamiento y fusión de dos o más palabras que forman un término complejo o sintagma (sobre todo por el principio de la primera y el final de la última), y que se pronuncian de forma silábica como una palabra.

Las siglas son abreviaturas de un término complejo al que sustituyen y están formadas por la yuxtaposición de las letras iniciales de las palabras que componen dicho término, pudiendo pronunciarse diciendo cada letra por separado o como una palabra completa.

Resulta interesante señalar que los términos simples pueden convertirse en términos complejos cuyo acrónimo forma parte, a su vez, de otro término complejo. Por ejemplo, el acrónimo "láser" se formó inicialmente mediante la fusión de las letras iniciales del término complejo original inglés "*l*ight *a*mplification by *s*timulated *e*mission of *r*adiation". Dicho término se convirtió en término simple y ahora forma parte de términos complejos tales como "rayo láser" o "arpa láser".

La terminografía se encarga de identificar y extraer de manera manual, generalmente[10], estas unidades simples o compuestas y su propósito principal es que se puedan organizar de manera sistematizada y armónica. Es, en pocas palabras, la rama que se encarga de aplicar los conocimientos que la terminología nos proporciona en pro de la obtención de productos como glosarios, vocabularios, diccionarios terminológicos y bases de conocimiento léxico entre otros. Puede asociarse a la práctica lexicográfica, derivada en gran parte de la lexicología, pero con un matiz especializado.

5.1.3. Terminótica

Ahora, dentro de los estudios actuales se puede ver que la separación entre las diversas ramas de la lingüística, en algunos casos, es un poco difusa debido al creciente número de personas involucradas en temas interdisciplinares además de la relación incipiente de casi todo el trabajo de investigación con los asuntos de la tecnología. Nos referimos aquí,

[10] Decimos "generalmente" ya que hasta hace algunos años las técnicas de extracción de términos divergían del trabajo que hoy se hace y, particularmente, del que se aborda en esta tesis. Más adelante se muestra porqué la extracción se considera "manual" al abordar el concepto *terminótica*.

específicamente, al tratamiento de la información especializada y su agilización en los procesos de extracción de información relevante vía procesos computacionales.

Desde hace poco más de dos décadas, la computadora ha venido a ser una herramienta indispensable en los estudios de cualquier tipo; así, la lingüística no escapa a esta interrelación humano-máquina e incluye toda serie de procesos lingüísticos asistidos o totalmente automatizados que vienen a darnos una visión más amplia y profunda acerca de los fenómenos tanto diacrónicos como sincrónicos. De manera particular podemos ver esta inmersión en la serie de diccionarios electrónicos que cada vez son más utilizados o preferidos por los usuarios que aquellos de una tradición parcialmente clásica -los que se publican en formato papel-. Si bien varios de los recursos lexicográficos que están en Internet son versiones electrónicas de los ya publicados anteriormente en papel (DRAE, VOX, etc.), algunos de ellos sólo los podemos encontrar en Internet, por ejemplo el Wordreference[11].

Partiendo de esta premisa, saber qué tipo de recursos son éstos últimos es importante ya que si bien los diccionarios electrónicos podrían considerarse parte de la lexicografía computacional, también los hay de lengua especializada -los terminológicos- que podrían insertarse en una terminología computacional o como se ha llamado actualmente, terminótica. Esta palabra es una fusión entre la práctica terminológica (terminografía) y la informática. Recordemos que la informática es "la ciencia aplicada que abarca el estudio y aplicación del tratamiento automático de la información, utilizando dispositivos electrónicos y sistemas computacionales" y también está definida como "el procesamiento automático de la información."[11]. Por lo tanto, es el recurso ideal para procesar la inmensa cantidad de información (el corpus) que se utiliza para crear un producto terminológico.

Con lo anterior queremos decir que describimos aquí parte del trabajo terminótico: la técnica y la metodología que utilizamos para procesar con herramientas computacionales todos aquellos datos del quehacer terminográfico, desde el proceso que

[10] http://www.wordreference.com/es/

[11] http://es.wikipedia.org/wiki/Informática

va de la constitución de nuestro corpus de trabajo, pasando por la metodología para la extracción terminológica, hasta llegar a la evaluación de los resultados obtenidos.

Más tarde, en el proyecto lexicográfico del GIL, nuestra meta es llegar hasta la extracción automática de definiciones y la consulta de los resultados obtenidos que no mencionamos en esta tesis pero que están planeados como trabajo futuro; estos procesos también están incluidos en la práctica de la terminótica.

Así, podemos ver que esta rama no sólo incluye la parte terminológica de nuestro estudio –ya que va más allá de la teoría- sino también la práctica terminográfica aunada a herramientas computacionales imprescindibles en nuestro tiempo para la creación de productos terminológicos ya sea en soporte electrónico o papel.

5.1.4. Extracción terminológica

La extracción terminológica es el proceso mediante el cual se seleccionan de un texto o un corpus unidades lingüísticas que pueden constituir términos. En ella, intentamos descubrir los términos más relevantes sin conocerlos previamente. Se trata de que las unidades seleccionadas vía herramientas computacionales sean las mejores candidatas a constituir términos simples o compuestos. Las herramientas que se utilizan generalmente identifican unidades con un mayor porcentaje de aparición, unidades importantes por su constante combinación con otras palabras o, incluso, por su aparición en ciertas partes del corpus aunque sean mínimas.

La necesidad de extraer términos de un área de especialidad es una tarea de creciente interés en la lingüística computacional. Para ello, es necesario un análisis lingüístico detallado que realizan expertos en el área de la terminología clásica aunados a los esfuerzos de expertos en computación, quienes diseñan herramientas para el óptimo desarrollo de esos análisis debido al avance de la tecnología. Al mismo tiempo, la metodología con que son superados los escollos que surgen en el proceso de recopilación del corpus y de la extracción de términos debe ser documentada en pro del avance de la lexicografía computacional para futuros estudios.

La extracción de términos a partir de diversas fuentes es la parte más importante de cualquier estudio terminológico, ya que de ella depende poder contar con el material

de análisis necesario. Para cualquier área de trabajo en la que se desarrolla un especialista es imprescindible tener a la mano las palabras que conforman el campo semántico de su área de estudio. Como ya mencionamos arriba, ese conjunto de palabras, llamados términos, le darán al especialista una herramienta inestimable de consulta para su quehacer científico, lo que impactará en la calidad de su trabajo y le ahorrará tiempo de consulta.

Aunque la extracción de términos es una tarea que se desarrolla en el área de lingüística principalmente, los productos derivados de esta son en su mayoría dirigidos a áreas que están fuera de ella. Por ejemplo, los sistemas para hacer minería de textos en áreas como bioinformática (Ananiadou y McNaught 2006) o medicina (Pustejovsky et al. 2002), hasta la construcción automática de ontologías para diversos fines (Buitelaar, Cimiano y Magnini 2008; Sierra et. al 2007).

La extracción terminológica se vale de conocimientos y técnicas de trabajo –como los de la ingeniería lingüística- que agilizan su desempeño como área de apoyo de cualquier campo de conocimiento. Es, pues, de suma importancia poder realizar una extracción con una metodología específica que garantice la calidad y la productividad de sus productos.

En el siguiente apartado describiremos de manera detallada cada uno de los pasos que seguimos para obtener la terminología básica en sexualidad para el español de México desde el corpus que conformamos y que documentamos en el capítulo 3 del presente trabajo.

En esta etapa final, lo que se trató de extraer fueron esas palabras utilizadas en el área de especialidad que aunque son conocidas por la comunidad que las utiliza no están indizadas de manera sistemática, ya que su clasificación y agrupamiento dependen de análisis detallados que se realizan gracias a la tarea constante del terminólogo. La meta es que cualquier persona pueda consultar el uso de esa palabra-término dentro de los diversos contextos del área de especialidad o dentro de los contextos fuera del área especializada con la mayor efectividad comunicativa posible.

5.1.5. Metodología para la extracción de términos

En sentido estricto, según el DRAE, una metodología es "el estudio del conjunto de métodos para llevar a cabo una tarea." Es decir, el trabajo contenido en una actividad para llegar a una meta es conocido como método, mientras que la metodología es el estudio de ese método o de varios de ellos si es el caso.

En nuestro caso, la metodología para la extracción de términos no ha sido desarrollada en su totalidad en México o no ha sido publicada debido a la reciente integración de esta área al campo de la lingüística computacional y debido a la juventud de la terminología como rama de la lingüística en nuestro país. Así, no sólo se busca aportar una aportación a la definición de *metodología* en el campo de la extracción de términos, también se busca que esa metodología propuesta se acerque lo más posible a la forma de trabajo de los expertos en la terminología en nuestros tiempos.

La extracción de términos es una tarea compleja ya que supone el análisis no sólo de todas las palabras que se insertan en el corpus de trabajo sino también de su relación con otras palabras. Como la metodología aquí desarrollada se ha basado en herramientas lexicográficas computacionales, tenemos que mencionar que para la obtención de la lista general de palabras, así como la de los candidatos a términos, utilizamos la herramienta *Wordsmith*[12], que es un programa computacional que nos guía a través de la generación automática de dichas listas con la ayuda de parámetros definidos por nosotros (por ejemplo, la supresión de palabras vacías o funcionales) y con el conteo basado en estadística.

Nuestra metodología, así planteada, supone todo el trabajo que va desde la conformación del corpus de trabajo hasta el diseño y la creación de la lista de términos que presentaremos como cumbre de este trabajo. La metodología, por lo tanto, es el trabajo terminótico en sí que lleva a cabo el Grupo de Ingeniería Lingüística de la UNAM con el fin de obtener los productos que se deriven del Corpus de las Sexualidades en México, en un principio, y de aquellos productos que se puedan obtener con dichas

12 Para mayor información sobre este programa se puede visitar la página: http://www.lexically.net/wordsmith/

herramientas y procesos a partir de los otros corpus con que cuenta nuestro grupo de investigación. A grandes rasgos la podríamos resumir de la siguiente manera:

1. *Diseño y recopilación de corpus*: Abarca la fase del nacimiento del corpus de trabajo.
2. *Limpiado y etiquetado de textos*: comprende la manipulación de grandes cantidades de texto para su posterior explotación.
3. *Explotación de corpus*: Incluye la exposición del corpus a diversas herramientas lingüísticas con el fin de extraer información específica.
4. *Análisis terminológico manual*: Es la fase donde se analizan los resultados arrojados por las herramientas terminóticas con el fin de establecer si estos son candidatos idóneos para crear productos terminológicos.
5. *Creación de extractores terminológicos automáticos*: Este punto, aunque no lo abarcamos en este trabajo, también se ha llevado a cabo en el GIL y consiste en aplicar algoritmos computacionales de agrupación y selección de candidatos a términos.[13]
6. *Evaluación de productos terminológicos*: Consiste en determinar si los productos derivados de la extracción terminológica derivados de esta metodología cumplen con su función de reflejar el estado actual del léxico especializado en el área para la cual fueron diseñados.

Hay que dejar claro, empero, que al decir metodología para la extracción de términos no se iguala a extracción terminológica. Mientras la primera trata de describir todo el trabajo para encontrar dichas unidades del lenguaje especializado, la segunda sólo desea documentar el proceso que va desde la identificación de palabras y sus paradigmas hasta su establecimiento como unidades especializadas en alguna área en específico.

En el siguiente aparatado veremos la parte de la extracción terminológica en el GIL después de que ya contamos con un corpus limpio y etiquetado, nada más. Con esto, esperamos dejar claros ambos conceptos: metodología para la extracción de términos y extracción terminológica.

5.2. La extracción terminológica en el GIL

La primera la distinción que hacemos dentro de la extracción terminológica es la que se da entre *token* y *type*. Mientras el primer término se refiere a cada una de las

[13] Para mayor información se puede consultar el artículo de Barrón A., Sierra G., Drouin P. Ananiadou S. de 2009 "An Improved Automatic Term Recognition Method for Spanish" en *Lecture Notes in Computer Science*. Número 5449. Springer-Verlag, pp. 125-136.

formas que aparecen en el texto, sin importar cuántas veces ocurra cada una (es decir, todos los grafos entre dos espacios en blanco); el segundo se refiere a cada una de las formas diferentes que aparecen en un texto. Así, el *type* agrupará dentro de sí todas aquellas palabras que se puedan unir por medio de un conjunto de grafos iniciales sin importar su origen flexivo o derivativo (como el conjunto *com-er, com-ida, com-ensal*), empero no aquellas que por su origen pertenecen a un mismo paradigma semántico pero morfológicamente distinto (como en el caso de *ir, voy, venga,* etc.)[14]. De tal manera que podemos entender que cada *token* será, en este primer acercamiento, igual a cada palabra existente en el corpus de trabajo. En breve, la suma de las frecuencias de todos los *types* será igual a la suma de de las frecuencias de todos los *tokens* de un corpus.

Así, la extracción de términos parte de esta distinción –recordemos que la computadora y las herramientas que utilizamos no identifican "palabras" tal cual nosotros las entendemos-.

La extracción terminológica que hicimos contempla tres fases que fueron supervisadas por cada una de las personas encargadas de cada área de la sexualidad que les correspondió. Así, primero fueron extraídos los términos de cada área y después se integraron para formar la terminología básica. Las fases las detallamos a continuación.

5.2.1. Primera fase: lista general de palabras

La primera fase consiste en la creación de una lista general de palabras con el fin de identificar cuáles podrían ser candidatas a términos de acuerdo con su porcentaje de aparición en el corpus. Para este fin, fueron eliminadas todas aquellas palabras funcionales o de vacio semántico (Ullmann, 1965) como los nexos y las conjunciones, para dejar sólo aquellas *plenas* (sustantivos, adjetivos, verbos, etc.).

En nuestro caso, la herramienta utilizada, Wordsmith, a través de su *Wordlist* define el conjunto de palabras como el listado de los *types* o formas que se escriben diferente en el corpus y nos arroja la lista de acuerdo con dos tipos de frecuencia que puede encontrar en el conjunto de textos: la *frecuencia absoluta* y la *frecuencia relativa*. La frecuencia absoluta es el número de veces que se repite cada palabra en el corpus. La

[14] Este problema se resuelve, en esta investigación, mediante revisiones y análisis manuales en el conjunto de *tokens* morfológicamente distintos que tienen que ser asociados a un solo *type.*

frecuencia relativa es el número de veces que ocurre la palabra repetida en relación con el total de palabras en el corpus; su valor es igual a la frecuencia absoluta entre el número total de *tokens* (o la suma de las frecuencias absolutas).

Esta lista de palabras es el punto de partida del análisis terminológico ya que no podemos igualar una entrada o lema[15] a un término, pero debido a que nuestro método se basa en herramientas informáticas, hemos decidido que otorgaremos la designación *término*[16] a toda aquella palabra introducida en una búsqueda siguiendo a Pape & Jones (1988).

5.2.2. Segunda fase: lista de palabras-clave

La segunda fase atiende a la creación de una lista de palabras-clave con ayuda de la herramienta *keywords* -que es parte de la paquetería del mismo *Wordsmith*- y análisis manuales. Lo que significa que después de seleccionar los elementos más frecuentes de la lista general de palabras, se procede a elegir los elementos con más alto rendimiento arrojados por esta herramienta.

Al referirnos a los elementos con más alto rendimiento queremos decir que estas palabras son aquellas candidatas a términos que pueden reflejar información relevante en una búsqueda (cualidades semánticas determinadas en contextos específicos). Cuando cumplen con esta condición hemos dado en llamarlas palabras-clave (o *keywords*) siguiendo también a Pape y Jones (1988) y al DRAE. Este último apunta que una palabra clave es aquella "palabra significativa o informativa sobre el contenido de un documento, que se utiliza habitualmente para su localización y recuperación en una base de datos"[17].

---

[15] Utilizamos aquí la palabra lema como sinónimo de entrada (de diccionario). No confundir con su homógrafo en Recuperación de Información (lemma) cuyo significado sugiere una forma canónica del primer conjunto de fonemas que pueden utilizarse para los fenómenos de flexión, derivación y conjugación sin llegar a ser la raíz de dicha palabra (es decir, reduce todo un paradigma flexivo y se toma como representante de todas las variaciones morfológicas de la palabra posibles) que en todo caso aquí correspondería a lo que llamamos *type*.

[16] En esta parte tomamos la definición de *término* de estos autores debido al procedimiento computacional para tratar las unidades de trabajo en la extracción terminológica. Como podemos ver, dista un poco de la acepción lingüística sin por esto excluir a la segunda –muy al contrario, complementa nuestra visión de dicha palabra-.

[17] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=t%E9rmino

También encontramos en palabras de Sierra (1999) que "el término palabra-clave es usado para designar cualquier palabra relevante usada en una consulta sintagmática con propósitos de recuperación (de información)". Con estas tres definiciones en mente creamos una lista de palabras-clave que nos diera la pauta para elegir los términos que finalmente se utilizaron para el *Vocabulario básico de las sexualidades en México*. Así, tenemos que una palabra-clave para nosotros es el punto más cercano a un término sin ser aquella definitiva ni suficiente para la elección de este.

El rendimiento se mide de acuerdo con dos parámetros que nos dan una medida. En conjunto, nos dicen qué tan eficiente resultó la búsqueda de las palabras clave de acuerdo con los parámetros que definimos como terminólogos, es decir, cuáles de las palabras extraídas contienen información relevante como para ser consideradas términos o candidatos a términos y cuál de esa información fue ignorada –lo cual nos crea una parquedad en la lista final de términos-.

Los parámetros de los cuáles hablamos son precisión y recuerdo. La precisión es el porcentaje de extracciones realizadas que fueron correctas y el recuerdo es el porcentaje de palabras clave con información relevante para ser consideradas término que fueron correctamente identificadas. Esto es, la precisión es igual al número de palabras clave extraídas entre la suma de las palabras clave extraídas más el número de palabras no-clave que se extrajeron; el recuerdo es el número de palabras clave entre la suma de las palabras clave más el número de palabras clave no extraídas. El análisis completo sobre la efectividad de la herramienta *keywords* de *Wordsmith* lo podemos encontrar en un interesante artículo de Michael J. Giarlo (2005). Para fines prácticos, nosotros trabajamos sobre la lista que nos arrojó *keywords y* después de descartar por método manual las palabras que aparentemente eran clave nos dimos a la tarea de definir una lista de candidatos a término.

5.2.3. Obtención de la lista de candidatos a término

En Pape y Jones[18] encontramos que "una palabra-clave refleja los conceptos inherentes en una consulta". De tal manera que para poder determinar esta lista de candidatos a

[18] *Ibid*

términos, tuvimos que atender a los criterios de la terminología donde los términos son definidos por un proceso de designación. Siguiendo a Wright y Budin (1997: 337), "los sistemas conceptuales y, por lo tanto, las definiciones en terminografía están basados principalmente en la extensión y la intensión de los conceptos, lo que hace que podamos hablar de definiciones terminográficas extensionales y definiciones terminográficas intensionales." (García de Quesada, 2001). Como no hablaremos aquí sobre las definiciones de los términos, basta con que se mencione esta sencilla distinción que hace que la labor terminográfica se vea indiscutiblemente separada de la labor lexicográfica clásica y, por lo tanto, que el proceso de elección de los elementos a definir sea distinto. Todo este proceso fue llevado a cabo de forma manual.

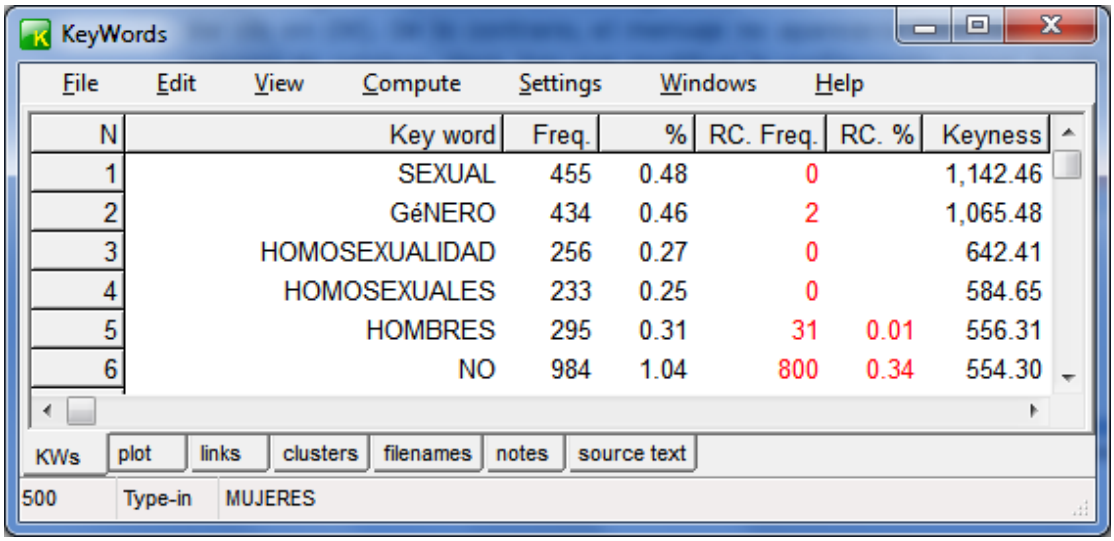
Como veremos a continuación la elección de un término incluye, además, análisis que tienen que ver con el uso y la distribución de una palabra-clave dentro del corpus de trabajo. Este uso y distribución son determinados, principalmente, por dos variables denominadas *peso* y *ranking*.

En recuperación de información (RI) se llama *asignación de peso* o *ponderación de palabra-clave* al proceso de asignar un peso –representado generalmente por un valor numérico– a cada elemento indexado de acuerdo con su importancia relativa (Sparck Jones & Kay, 1973: 146). Nosotros nos apoyamos para asignar estos pesos a los candidatos a términos de acuerdo con su posición en la lista de *keywords;* es decir, de acuerdo con su porcentaje de aparición en el área de la sexualidad a la que corresponden dentro del corpus de trabajo y, por la "rareza" de su aparición (ya que en consecuencia las palabras-clave menos frecuentes serán también más importantes) en la totalidad de los textos analizados.

El *ranking* o posicionamiento es usado en recuperación de información para describir la capacidad de un sistema para descartar las palabras irrelevantes de las que sí lo son y presenta en secuencia aquellas que son relevantes en orden de importancia con respecto a una búsqueda. Este posicionamiento sirve para determinar si un candidato a término cumple con algunas restricciones como la aparición junto a otras palabras que la definen o complementan (en el caso de los términos multipalabra) o si sólo se trata de un elemento del conjunto de una palabra altamente encontrada. Esto quiere decir que una palabra se puede encontrar en cualquier lugar del corpus de trabajo pero dependiendo de

su posición o cotexto y su situación (Adelstein, 2005), esta palabra será un término o un elemento más en el texto.

En primer lugar nos dimos a la tarea de seleccionar los términos simples, es decir los términos que se conforman por una sola palabra. Como podemos ver en la siguiente imagen, en la lista pueden aparecer tanto términos como palabras comunes. De esta lista se seleccionaron manualmente los términos de un área temática en particular, una por una. Las frecuencias que aparecen bajo la columna Freq. son las frecuencias absolutas de las palabras listadas. Así, el número de apariciones de la palabra, aunado a su posicionamiento y la discriminación manual nos da como resultado la lista de términos que buscamos.



Cuando hablamos de discriminación manual queremos decir que hicimos una revisión visual de la lista de términos seleccionados porque este programa aún no es tan efectivo como para confiar en él al cien por ciento. Queremos decir, por lo tanto, que hay ruido dentro de dicha lista (palabras vacías, símbolos, fórmulas, etc.)

Ahora, para hacer esta selección nos apoyamos en dos conceptos básicos: *discriminación de palabras-clave* e *indización probabilística*. El primero nos ayuda a identificar a qué área de la sexualidad corresponde una palabra-clave en contraste con la colección de textos completa. El segundo rasgo nos indica el "peso" de la palabra clave conforme a su ocurrencia en los dos polos de los textos de la colección: el más relevante y el menos relevante; esto con respecto a una determinada búsqueda. Pudimos distinguir las áreas a las que pertenecían los términos gracias a que previamente separamos los

textos de acuerdo con cada humano-procesador de los textos del corpus. Así, el encargado del área de *educación sexual* por ejemplo, hizo una extracción terminológica de evaluación antes de fusionar sus documentos con el resto del corpus. Una vez hecha la extracción final, se cotejó la lista de la terminología básica con cada una de las listas previas en cada área.

Recordemos que uno de los fines esenciales de nuestro estudio es que no exista un área más relevante que otra (por ser nuestro corpus de naturaleza equilibrada) y que aún así podamos encontrar una lista de palabras-clave que sean variadas y candidatas a términos. Así, establecer el peso de una palabra-clave tiene un fin primordial: especificar la importancia de dicha palabra dentro de una búsqueda y, por ende, determinar si esa palabra es un término o no.

Para los términos complejos o multipalabra utilizamos una herramienta dentro del Wordsmith conocida como *Concord*. Para ello, sobre la lista de palabras clave que generamos, hicimos una lista más con las palabras que se relacionaron con los candidatos a término al menos dos posiciones a la derecha y dos a la izquierda. De esta lista, elegimos los que determinamos como términos apropiados (por mencionar un caso, para el término "identidad" se seleccionaron los términos complejos "trastorno de identidad" y "trastorno de identidad sexual").

5.3. Resultados obtenidos

Al terminar la discriminación de los términos multipalabra la lista creció un poco y nos arrojó como resultado un total de 1285 términos entre simples y complejos, todos los cuales pueden ser consultados en el Apéndice B de esta tesis.

Finalmente, una vez realizados los análisis anteriores, en la última fase se procede a la creación de listas de términos según las ocho subáreas de la sexualidad establecidas arriba. Así que elegimos 125 términos por cada área con el fin de obtener un total de mil términos. Estos términos son los que definimos como terminología básica (ver Apéndice D).

A manera de resumen podemos decir que con una cantidad de textos como la que hemos recuperado es relativamente fácil obtener una lista armonizada de

aproximadamente 1285 términos limpios. Esto es, una terminología sin ruido que nos puede decir que es un trabajo básico y confiable.

En cuanto al tiempo de procesamiento pudimos ver que el proceso nos llevó aproximadamente nueve meses: seis meses en los cuales trabajamos como equipo para la recuperación, limpiado y procesamiento del corpus y tres meses de revisión, etiquetado, y codificación. Además, tenemos que agregar un periodo de tres meses más para la creación de la base de datos y la revisión de los documentos en codificación XML, ya que es el producto que finalmente se sube a Internet para que el generador de concordancias pueda trabajar. De tal manera que de acuerdo con el tiempo, una extracción terminológica de esta naturaleza se puede llevar a cabo en aproximadamente un año.

En cuanto a la confiabilidad podemos decir que es un trabajo eficaz que refleja con aceptable amplitud el léxico que el mexicano actual utiliza en el área de sexualidad. Los parámetros para decir esto se basan en la cantidad de términos que se extrajeron aunados a la cantidad de textos procesados.

Un resultado más es la nueva interfaz que se trabaja en el Grupo de Ingeniería Lingüística. Entre sus nuevas características se encuentra un motor más veloz para la generación de concordancias y un indexador mejorado que impide la pérdida de elementos lingüísticos dentro de la nueva versión del corpus. Esta interfaz se puede consultar en: http://www.iling.unam.mx:8080/csmx2.

Aunado a esto podemos encontrar la conclusión exitosa del proyecto *Proyectos de apoyo para investigadores nacionales para el fortalecimiento de actividades de tutoría y asesoría de estudiantes de nivel licenciatura (SNI-estudiantes)* con un total de cuatro publicaciones entre nacionales e internacionales que documentan el proceso de creación de la terminología básica de las sexualidades en México.

El último de los resultados es una lista de los productos y los beneficios obtenidos. Entre ellos podemos contar:

- Una nueva versión del *Corpus de las Sexualidades en México* (CSMX2).
- Una nueva versión de la interfaz para la visualización del CSMX y la mejora y actualización del generador de concordancias.

- Ampliación del tema de la terminótica en México a través del trabajo activo y la promoción de los resultados previos a esta tesis.
- Inclusión de nuevas personas al equipo de trabajo de corpus –en particular al de sexualidad- dentro del Grupo de Ingeniería Lingüística y, por extensión, al área de la Ingeniería Lingüística en México y en la UNAM.
- Una lista de términos con los que podemos experimentar para el mejoramiento y desarrollo final de la herramienta DESCRIBE®.

6. Conclusiones

En este último capítulo se presentan las conclusiones del presente proyecto de tesis. Se trata esencialmente de una revisión sobre la rapidez y efectividad en el diseño del CSMX, la cadena del procesamiento de los textos que conforman el CSMX y la calidad del generador de concordancias, aunados a la evaluación sobre la extracción semiautomática que hemos hecho para la creación de nuestra Terminología Básica de las Sexualidades en México.

En el capítulo 2 presentamos una descripción sobre dos términos que desde hace años se vienen trabajando en México pero que desgraciadamente aún están en una etapa temprana de desarrollo: el corpus lingüístico y la lingüística de corpus. En cuanto al primer término podemos mencionar que los criterios para su creación son variados e importantes, van más allá del análisis de la lengua sobre la cual el lingüista mexicano desarrolla su trabajo, es decir, al mismo tiempo que dicho lingüista trabaja sobre los diferentes fenómenos, también se tiene que preocupar de la creación de las herramientas que lo harán llevar sus investigaciones a buen puerto.

Si bien los trabajos actuales en México son de una calidad notable en esta área de las humanidades, los especialistas dedicados a ella no han impreso un sello distintivo en la creación de corpus lingüísticos en soporte electrónico -sólo por mencionar el tipo de corpus que nosotros construimos, ya que como bien se menciona, no es el único formato en el cual podemos encontrarlos-. Así, este capítulo muestra la revisión de los criterios que se deben acatar idealmente para la creación de dicho material. Con esta revisión llegamos al segundo término: el experto en el análisis de los criterios formales para la creación de corpus será un especialista en lingüística de corpus. La importancia de este capítulo radica justo en eso, mostramos que no es suficiente contar con una cantidad importante de textos para hacer un análisis lingüístico específico, sino que también es de suma importancia que este material esté organizado de acuerdo con el fin al cual va destinado y que su organización esté sistematizada para que la investigación lingüística sea más fina, eficaz y veloz al momento de buscarlo o intentar procesarlo. Mostramos, pues, que la lingüística de corpus sí existe en México y que hay trabajos sobre ella que nos llevan al diseño de corpus lingüísticos informatizados de una buena calidad.

El capítulo 3 nos muestra la fase de creación del Corpus de las Sexualidades en México (CSMX), abarca desde la revisión de los criterios que abordamos en el capítulo 2

y su aplicación a un trabajo práctico y desglosa cada uno de los apartados a los que sometimos la colección de textos para convertirlo en un corpus confiable, variable, equilibrado y representativo.

A través del ejercicio consciente sobre cada uno de los detalles que tiene la colección de textos, mostramos que la construcción de nuestro corpus fue disciplinada y revisada bajo los estándares más precisos posibles. Vimos que si bien no se puede llegar a una representatividad total de una lengua por más extenso o riguroso que sea un corpus, sí se puede llegar a contar con un material confiable y amplio en cuanto a la variedad léxica corresponde. Al final de este capítulo el lector puede caer en la cuenta de que se trata de una organización jerárquica y armonizada de textos que siempre tiene una base detrás: versiones originales, anotadas, codificadas en diversos lenguajes, con anclas a imágenes, otras páginas electrónicas, bibliografía complementaria, etc. y que es necesario y recomendable guardarlas todas porque cada una tiene un fin específico. Como mostramos al final del capítulo, nosotros decidimos contar con una copia del corpus sin anotación o codificación alguna (tenerlo en texto plano) ya que nuestras necesidades de procesamiento exigían contar con un corpus textual limpio.

Este capítulo muestra, pues, que antes de cualquier análisis lingüístico es necesario contar con un material óptimo, ya que de ello dependerán distintas variables que determinarán el éxito o alcance de nuestra investigación. El CSMX cuenta, como se ve aquí, con un bagaje teórico que nos permite dedicarle un capítulo completo en aras de reivindicar la lingüística de corpus como una rama necesaria de la lingüística en general en pro del avance y la exactitud investigadora de esta área.

En el capítulo 4 hacemos un recorrido por las herramientas que utilizamos para el procesamiento del corpus. Si bien nuestra tarea principal, la extracción terminológica, exigía que nuestro corpus estuviera limpio, también fue necesario contar con una versión anotada para que el corpus no quedara como un material de uso exclusivo del Grupo de Ingeniería Lingüística. De tal manera que fue necesario el uso de programas computacionales que nos ayudasen en dicha tarea con el fin de poner a disposición del público en general todos aquellos textos que fueron recuperados de Internet a través de un generador de concordancias que finalmente mostramos como la interfaz de consulta.

Ahora bien, es importante señalar que la mayoría de las herramientas que mencionamos en este capítulo hacen referencia al trabajo que se ha hecho en el GIL desde hace poco más de una década, pero el recorrido que hacemos no sólo intenta mostrar que dichas herramientas nos ayudan en la cadena de procesamiento y nos ahorran días o meses de trabajo manual, sino también intenta mostrar que la ingeniería lingüística actúa en beneficio de todo lingüista interesado en ahorrar tiempo al trabajar con grandes cantidades de texto. Estas herramientas son sólo una prueba de que una de las áreas en la cual se inserta este trabajo –la ingeniería lingüística- nos ayuda de manera inconmensurable en la labor lingüística, ya sea que trabajemos en terminología, lexicografía, análisis del discurso, sintaxis, etc. Es, como mencionamos continuamente, el área que se encarga de resolver problemas lingüísticos con la ayuda de la computadora y herramientas informáticas y, por lo tanto, es realmente necesario mostrar cómo es que su labor es ya imprescindible en cualquier trabajo que llevemos a cabo, pues ¿qué lingüista no ocupa en estos días la computadora, qué investigador? Los programas que mostramos son apenas la punta de un iceberg de posibilidades que ya existen y que, desgraciadamente, pocos utilizamos al momento de hacer lingüística teórica o aplicada. Este capítulo muestra que de a poco cada investigación tendría que contener al menos una descripción somera sobre las herramientas que se utilizan para el desarrollo de dichas investigaciones; desde el procesador de textos hasta, por ejemplo, un analizador automático de sentimientos.

En el capítulo 5 describimos, en principio, las bases teóricas sobre las que sentamos la extracción terminológica semiautomática que llevamos a cabo. En esta primera parte nos enfocamos en mostrar que lo que hemos hecho para crear la terminología básica no es ni un trabajo exclusivo de la terminología ni mucho menos un producto más de algún programa computacional. Más allá de lo que se puede observar a simple vista, la extracción terminológica es, como lo desarrollamos, una interacción entre varias ramas de la lingüística (terminología, terminografía, lexicografía, terminótica, etc.) aunada a varias ramas de cómputo (extracción de información, minería de textos, etc.) que en conjunto nos ofrecen un producto novedoso (pues no existe una terminología básica en sexualidad para el español de México) y que satisface una necesidad lingüística en nuestro país. A través de la descripción de las tres fases en las cuales desarrollamos el

producto final, mostramos la importancia de la interacción entre computólogos y lingüistas en pro de la creación de materiales y herramientas útiles y necesarias en ambos campos.

Como trabajo futuro, la terminología creada se utilizará en conjunto con el sistema Describe®, "un sistema que permite obtener de manera resumida y organizada la descripción completa de un término" (Molina, 2008) para obtener las definiciones de los mil términos obtenidos. Este sistema, en fase de prototipo, está siendo desarrollado por el Grupo de Ingeniería Lingüística en la Universidad Nacional Autónoma de México. De tal manera que establecida la terminología básica, se procederá a la búsqueda de definiciones de esos términos de manera automática y semiautomática. La herramienta Describe® busca ser una ayuda indispensable en la conformación de definiciones de términos a partir de la búsqueda de éstas en la Web. Esto tiene dos ventajas principales: en primer lugar el lexicógrafo verá en ella un apoyo excelente al momento de tratar de integrar todas aquellas palabras que deberían estar en su definición, y por otro lado, este mismo especialista tendrá la oportunidad de contar con diversas variantes semánticas de una palabra en el mismo momento de su cambio. Es decir, a través de la búsqueda en Internet esta herramienta y este método pueden integrar al conocimiento del lexicógrafo información actualizada día por día con una sola ejecución del programa. En estudios sincrónicos sobre la lengua es de suma importancia contar con esta información ya que nosotros, como hablantes de una determinada lengua, sentimos como única la definición que tenemos en nuestro léxico activo y esa definición no atiende de primera mano las variaciones diacrónicas de la palabra en juego. Así, un especialista consciente de este fenómeno, podrá llevar a cabo un análisis más fino y minucioso sobre el cambio semántico de las palabras que intenta definir en su producto lexicográfico.

Al final de esta tesis podemos hacer un recuento de los beneficios y productos derivados de esta investigación a la par de mostrar un camino eficaz, veloz y altamente productivo en el área de la terminótica. La conclusión de esta tesis nos mostró que es posible, en estos días, crear productos como corpus y diccionarios especializados de la calidad de las grandes editoriales en un tiempo mucho menor si hacemos coincidir la terminología y la ingeniería en computación. Apenas con esta muestra del gran trabajo que puede abarca la ingeniería lingüística nos damos cuenta de que el futuro de la

lexicografía y la terminología en México puede tener un horizonte más amplio gracias a las tecnologías del lenguaje.

Bibliografía

Adelstein, A. (2005) "La información semántica especializada de unidades léxicas simples" en *Traducción y Terminología. Entre teoría y práctica*. Facultad de Filosofía y Letras, Universidad Nacional de Tucumán, Argentina.

Ananiadou, S. & McNaught, J. eds. (2006) *Text Mining for Biology and Biomedicine*. Artech House.

Boas, F. (1940) *Race, language and culture*. Chicago, The University of Chicago Press.

Buitelaar, P; Cimiano, P. & Magnini, B. eds. (2008) "Ontology Learning from Text: An Overview" en *Ontology Learning from Text: Methods, Evaluation and Applications Frontiers* en Artificial Intelligence and Applications Series, Vol. 123, IOS Press.

Cabré, M. T. (1998) *La Terminología. Teoría, metodología y aplicaciones*. Barcelona, Atàrtida-Enpúries.

Cabré, M. T. (1999) *La terminología: Representación y comunicación.* Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Cabré, M. T. (1999b) "Una nueva teoría de la terminología: de la denominación a la comunicación" en *La terminología. Representación y comunicación*, Barcelona, Institut Universitari de lingüística Aplicada, Universitat Pompeu Fabra.

Caravedo, R. (1999) *Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español.* Salamanca: Ediciones Universidad de Salamanca.

Fellbaum, C. ed. (1998) *WORDNET. An electronic lexical database.* Massachusets: The MIT press.

García de Quesada, M. (2001) "Estructura definicional terminográfica en el subdominio de la oncología clínica" en *Estudios de Lingüística del Español*, Vol. 14. UAB, Barcelona.

Garduño, G., Sierra, G., & Medina, A. (2004) "Herramientas de análisis para el Corpus Lingüístico en Ingeniería" en *Avances en las Ciencias de la Computación*, Instituto Politécnico Nacional, México.

Garrido Almiñana, J. M. (1996) *Modelling spanish intonation or text-to-speech applications.* Universitat Autònoma de Barcelona, Departament de Filología Espanyola. Barcelona: PhD Thesis.

Giarlo, M. (2005) *A Comparative Analysis of Keyword Extraction Techniques*. Rutgers, The State University of New Jersey.

Kilgarriff, A. & Grefenstette, G. (2003) "Web as a Corpus" en *Computational Linguistics*. Volume V, Number 29.

Leech, G. (1997) "Introducing Corpus Annotation" en Roger Garside et al. *Corpus Annotation*. Longman, London.

McEnery, T., & Wilson, A. (1996) *Corpus Linguistics.* Edinburgo: Edinburgh University Press.

Medina A., Sierra G. (2004) "Criteria for the Construction of a Corpus for a Mexican Spanish Dictionary of Sexuality" en *Proceedings 11th Euralex International Congress*. Vol. 2. Université de Bretagne-Sud. Lorient, Francia.

Medina, A., & Méndez, C. (2006) "Arquitectura del Corpus Histórico del Español de México" en *Avances en Ciencias de la computación*, Instituto Politécnico Nacional, México.

Molina, A. (2008) "Deja de buscar, mejor DESCRIBE" en *Actas del 4to Seminario de Ingeniería Lingüística*. Instituto de Ingeniería, UNAM. México, D.F.

Pape D.L. & Jones, R.L. (1988) "STATUS with IQ-escaping from the boolean stratjacket" en *Program* 22 (1).

Pavel, S. & Nolet, D. (2002) *Manual de terminología*. Ministro de Obras Públicas y Servicios Gubernamentales de Canadá. Canadá.

Pineda, L. A. (2009) *The Corpus DIMEx100: Transcription and evaluation*. Netherlands: Springer.

Procházková, P. (2006) *Fundamentos de la lingúística de corpus: Concepción de los corpus y métodos de investigación con corpus* en http://www.prochazkova.de/fundamentos_de_la_linguistica_de_corpus.pdf . Consultada en marzo 10, 2009.

Pustejovsky, J. et al. (2002) "Anaphora resolution in biomedical literature" en *International Symposium on Reference Resolution in NLP*. Alicante, España.

Rondeau, G. (1984) *Introduction à la terminilogie*. Québec, Gaëtan Morin.

Sánchez, A. (2008) *John Sinclair (1933-2007). In memoriam* en *Asociación Española de Estudios Anglo-Norteamericanos*: http://www.aedean.org/NEXUS-Archive/Nexus2008.1/Nexus%202008.1-94.pdf

Sánchez-Cuadrado, S. (2007) "Definición de una metodología para la construcción de Sistemas de Organización del Conocimiento a partir de un corpus documental en Lenguaje Natural" en *Procesamiento del Lenguaje Natural* nº39, SEPLN, España.

Sierra, G. (1999) *Design of a concept-oriented tool for terminology*. UMIST, Manchester, UK. PhD Thesis.

Sierra, G. (2006) "Diseño de corpus textuales para fines lingüísticos" en *Memorias del IX Encuentro internacional de Lingüística del Noroeste. 2*, pp. Sonora: Unison.

Sierra, G., & Rosas, A. (2008) "Clasificación de corpus textuales" en *Actas del X Congreso de Lingúística del Noroeste* (p. en prensa). Sonora: Unison.

Sierra, G.; Medina, A.; Lázaro, J. (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*. Mayo de 2009. AELINCO, Murcia, España.

Spärck Jones, K. & Kay, M. (1973) *Linguistics and Information Science*. Academic Press, New York.

Toruella, J. & Llisterri, J. (1999) "Diseño de corpus textuales y orales" en J. M. Blecua, G. Clavería, C. Sánchez y J. Toruella, *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Milenio

Ullmann, S. (1965) *Semántica: Introducción a la ciencia del significado*. Aguilar, Madrid.

Wright, S. & Budin, G. eds. (1997) *Handbook of Terminology Management*. Vol. 1. John Benjamins, Amsterdam.

Apéndice A

Fuentes del CSMX

ÁREA	TITULO ARTÍCULO	AUTOR	NIVEL	NOMBRE_REVISTA	NUM_PUBLICACION	FECHA_DE_CAPTURA	URL	CLAZ
comportamiento	IDENTIDAD DE GÉNERO	CARMEN RAMOS ESCANDÓN	3	LA VENTANA	10	09-oct-07	http://www.publicaciones.cucsh.udg.mx/pperiod/laventan/Ventana10/ventana10-7Carmen.pdf	comportamiento pdf 1
comportamiento	El mito del orgasmo vaginal	Anne Koedt	1			10-oct-07	http://www.equidad.org.mx/ddeser/seminario/internas/lecturas/lect-sexual/elmitodelorgasmovaginal.pdf	comportamiento academico 2
comportamiento	Relaciones sexuales durante el embarazo	M. en C. José Jaime Martínez Salgado	4			11-oct-07	http://www.consultasexual.com.mx/Documentos/embarazo.htm	comportamiento HTML 3
comportamiento	Ella encima de él		4			12-oct-07	http://www.terra.com.mx/articulo.aspx?ref=0&articuloid=129189	comportamiento HTML 2
comportamiento	Cuerpo a Cuerpo ¿Se puede tener relaciones durante el embarazo?	PEDRO AVALOS HUIZAR	4			13-oct-07	http://www.lajornadajalisco.com.mx/2007/03/03/index.php?section=sociedad&article=008a1soc	comportamiento HTML 1
comportamiento	¿Se puede elegir el sexo del bebé?	Jorge Tirado Chávez, Cirujano Ginecoobstetra y Psicólogo	4			14-oct-07	http://www.bbmundo.com/bbfertilidad/leerNota.asp?idSub=1&idArt=1855	comportamiento HTML 4
comportamiento	Alumnos tomaban fotos "porno" con celular: AEPAF	desconocido	3			15-oct-07	http://www.ehui.com/pdf.php?a=64801	comportamiento pdf 2
comportamiento	Sexualidad y género: algunas experiencias de investigación en México'	Ivonne Szasz	3			16-oct-07	www.equidad.org.mx/ddeser/seminario/internas/lecturas/lect-sexual/sexualidadygenro.pdf	comportamiento pdf 3
comportamiento	LAS MEJORES POSICIONES PARA UN PLACER AAL ORGASMO		5			17-oct-07	http://groups.google.com.mx/group/el-sexo-y-la-pareja/browse_thread/thread/3a0a2902f57f6fe7/7ed24eb07d6d97c7?lnk=st&q=posiciones+en+el+acto+sexual&rnum=24&hl=es&_done=%2Fgroup%2Fel-sexo-y-la-pareja%2Fbrowse_thread%2Fthread%2F3a0a2902f57f6fe7%2F7ed24eb07d6d97c7%3Flnk%3Dst%26q%3Dposiciones%2Ben%2Bel%2Bacto%2Bsexual%26rnum%3D24%26hl%3Des	
comportamiento	Como volver loco en la cama a cada signo					18-oct-07	http://groups.google.com.mx/group/korin/browse_thread/thread/974c0936a054a8e8/5dae76ee7e0f15b9?hl=es&lnk=st&q=posiciones+sexuales&rnum=2#5dae76ee7e0f15b9	
comportamiento	conocimientos objetivo y subjetivos sobre el VIH-SIDA como predictos de uso de condon en adolecentes	Alberto Villaseñor-Sierra, Dr CM,(1) Ramiro Caballero-Hoyos, Dr en CS,(2) Alfredo Hidalgo-San Martín, MSP,(2) José Ignacio Santos-Preciado, Dr CM,(3,4)	1			19-oct-07	http://scielo.unam.mx/scielo.php?script=sci_pdf&pid=S0036-36342003000700010&lng=es&nrm=iso&tlng=	comportamiento academico 1
comportamiento	Estrategias novedosas de prevención de embarazo e ITS/VIH/sida entre adolescentes escolarizados mexicanos	Pilar Torres, Antrop[I]; Dilys M. Walker, Esp en Ginecol y Obst[II]; Juan Pablo Gutiérrez, MC[III]; Stefano M. Bertozzi, Dr en Gerencia y Políticas de Salud[III, IV]	1			20-oct-07	http://scielo.unam.mx/scielo.php?script=sci_arttext&pid=S0036-36342006000400005&lng=es&nrm=iso&tlng=es	
comportamiento	FANTASÍAS SEXUALES, SOY LO QUE QUIERO SER		2			21-oct-07	http://www.saludymedicinas.com.mx/nota.asp?id=775	comportamiento asociaciones 3
comportamiento	Posiciones Sexuales		5			22-oct-07	http://rockandroll.foromup.com.mx/about154-rockandroll.html	
comportamiento	Hablando de Sexualidad Erotismo		2			23-oct-07	http://www.geishad.org.mx/erotismo.html	comportamiento asociaciones 1
comportamiento	Es un error tratar de imitar a un Casanova sexual		2			24-oct-07	http://www.paraqueestesbien.com/hombre/torax/saludhombre/salud33.htm	comportamiento asociaciones 2
comportamiento	¿Es frecuente la disfunción eréctil?		2			25-oct-07	http://medicaangelus.com.mx/Preguntas%20Sobre%20Disfuncion%20Erectil.htm	comportamiento asociaciones 4
comportamiento	Posiciones sexuales		5			26-oct-07	http://groups.google.com.mx/group/Drporn/browse_thread/thread/53a2c321e2ad209d/f8d5bb13e334da6b?hl=es&lnk=st&q=posiciones+sexuales#f8d5bb13e334da6b	
comportamiento	Kama Sutra		5		72	27-oct-07	http://groups.google.com.mx/group/cosasdivertidas/browse_thread/thread/1fc27b7e7bd18558/e047b4b093fc4136?hl=es&lnk=gst&q=posiciones+sexuales#e047b4b093fc4136	comportamiento foros 1
comportamiento	Diversidad sexual, un ejemplo de intolerancia	Alejandra Tello	3			28-oct-07	http://www.comsoc.udg.mx/gaceta/paginas/256/256-1415.pdf	
atraccion	Trivialis: erotismo, pornografía y obsenidad en la	Fabian Giménez	académico	*Discurso visual. Revista*	enero-abril 2006	octubre 7- 2007	www.discursovisual.cenart.gob.mx/anteriores/dvwebne05/agora/agofabian.htm	atraccion academico 1

Apéndice B

Términos simples agrupados por área

Expresión			**Fundamentos**	**Atracción**		**Educación**	**Sexualidad variante**	**Fundamentos**	**Educación**	**Respuesta**
sexual	sexo	no	aborto	Adolescencia	aborto	Sexual*	Sexo	actividades	clímax	sexualidad
homosexualidad	placeres	sexualidad	actividad	Afectiva	abrazos	sexualmente	Sexualidad	Alargamiento	amorosa	sexual
homosexual	natural	más	acto	Amor	abstinencia	Sexualidad	Vagina	amigos	reproductor	sexo
sexualidad	cuerpo	amor	adrenalina	Antropología	adolescencia		Salud	Anatom	familia	salud
menopausia	femenina	sexual	amor	Arquitectura	amante	Mujer*	Pene	Andropausia	prostitutas	mujer
SEXO	cerebro	ser	anatómicas	Art	amor	Salud	Orgasmo	animal	infección	amor
sexuales	explorar	padres	anticonceptivos	Artes	anal	Sexo	Sexología	animales	sadismo	persona
mujer	vivencia	sus	aparato	Asexual	anticonceptivos	Educaci*	Anal	ano	menstruación	culo
hijos	órgano	si	besos	Autoconocimiento	atracción	Hombre*	Culo	Anticoncepci	himen	pareja
hombre	opiniones	vida	biológicos	Bienestar	besos	SIDA	Sensual	Anunciar	lesbianas	hombre
relaciones	reproducción	sobre	caricias	Calor	bisexual	VIH-SIDA	Pareja	arriba	homosexualidad	vida
padres	largo	esta	cerebro	Cambio	cachondo	AIDS	Cunnilingus	atracci	anticoncepción	derecho
genes	estudio	cuando	coito	Ciencias	caliente	Pareja*	Anilingus	axilas	padres	adolescente
heterosexual	macho	sexo	componentes	Comportamiento	cama	emparejarse	Infidelidad	boca	herpes	problema
hablar	información	hay	comportamiento	Crisis	caricias	Enfermedad*	Lesbianas	bulo	vigor	humano
Anticonceptivos	parto	puede	condón	Cuerpos	castración	relaci*	Homosexuales	Circuncisi	afrodisíacos	cancer
cuerpo	mujeres	también	conquista	Culo	clítoris	Abort*	Heterosexual	columna	psicosexual	adolescencia
sexólogos	incontinencia	todo	conservación	Cultura	condones	Diversidad	Penetración	Comportamiento	autoestima	Reproductiva
vida	curso	pero	creencia	Cursos	contagio	Informaci*	Kamasutra	comprende	pubertad	abuso
cambios	aburrido	tiene	cuerpo	Desarrollo	cuerpo	Embaraz*	Felación	comunes	educación	Mostrar
desarrollo	adolescencia	hombre	cultural	Descendencia	depravación	Vida	Pornografía	Condones	anticonceptivos	relación
diferencias	amor	sexuales	dependencia	Deseo	deseo	Reproduc*	Placer	contactos	responsabilidad	disfunción
edad	celos	persona	deseo	Edad	desnudo	Anticonce*	Excitación	Copyright	Menopausia	edad
heterosexualidad	Charlas	hijos	disfunción	Educativo	dolor	Porno*	Deseo	costillas	Fertilidad	psicología
responsabilidad	Comunidad	educación	disfrutar	Embarazo	embarazo	Envejec*	Clímax	cuerpo	Parto	pene
chicas	Consultar	esto	elección	Enfermedad	emociones	Homosexual*	Juegos	cuyo	Hormonales	sensualidad
conducta	coqueteo	pareja	embarazo	Erecciones	enamorado	homosexualidad	Lubricación	Disfunci	Sexos	femenina
crianza	cuerpo	otro	embrión	Erotismo	espermatozoides	Viol*	Frigidez	educativos	ternura	trastornos
embarazo	culos	relación	entorno	Especies	esterilidad	Menopausia	Sida	ejemplo	edad	información
humana	Curso	entre	erección	Estimulante	excitado	Sexolog*	Masturbación	Enfermedad	riesgo	transmisión
placer	divertido	está	eróticas	Existencia	eyacular	Vagina*	Mitos	enfermedades	familia	adulto
valores	energía	uno	especie	Fecundar	falo	derechos	Gay	entre	virginidad	embarazo
abuso	esperma	cada	estrógenos	Frigidez	gonorreas	Gay*	Culpa	ER	satisfacción	Excitación
adolescencia	estudio	familia	espermatozoide	Gemelos	herpes	erección	Disfunción	Er	estimulante	gusta
adultos	familia	mujer	excitación	Hembras	Heterosexual		Erección	er	penetración	padres
cuerpos	femenina	bien	existencia	Heterosexuales	heterosexuales	Fantasía*	Intimidad	escroto	seropositividad	SIDA
Expresión			**Fundamentos**	**Atracción**		**Educación**	**Sexualidad variante**	**Fundamentos**	**Educación**	**Respuesta**
etapas	garantizada	dios	fantasía	Historia	ingles	Lesb*	Relaciones	estimulaci	Orgasmo	deseo
familia	Gratis	forma	fases	Hombre	jóvenes	Adolece*	Psiquiatra	Excitaci	Coito	Sexología
femenina	hijos	mismo	fecundación	Homosexualidad	juegos sexuales	Joven*	Freud	excitaci	Responsable	enfermedad
infancia	hombre	muy	físico	Humana	lesbianas	VIH	Amor	Eyaculaci	Paternidad	Esconder
matrimonio	humano	desde	fisiológica	Humanidades	lujuria	Orgasmo	Genitales	eyaculaci	masculina/ masculino	homosexual
peligros	importancia	sí	funciones	Identidad	machismo	conducta/s	Órganos	Femenina	prevención/ prevenir	Obesidad
responsables	Información	sólo	geneticamente	Individuo	marido	disfunci*	Comportamiento	Gay	juegos	placer
salud	Inicio	siempre	generación	Infancia	menstruación	amor	Coito	Gays	libido	prostitución
sentimientos	Instituto	así	genital	Jalea	miembro	Masturba*	Fantasía	GENA	mitos	experiencia
varones	Medicina	placer	gestación	Joven	morbo	Femenin*	Estimulación	gena	preservativo	HISTORIA
menstruación	menopausia	debe	goce	Lenguaje	nalgas	vejez	Disfunción	GENAS	Fantasías	orgasmo
adolescentes	menstruación	relaciones	hombre	Lesbiana	novia	cuerpo	Trastornos	genas	Enamoramiento	comportamiento
amistades	mundo	hacer	homosexualidad	Libertad	noviazgo	corporal	Satisfacción	genitales	Enamorados	erección
amor	Noticias	tener	hormonas	Libido	orgasmo	Píldora*	Anatomía	Hipospadias	Disfunción	gente
bisexual	novia	nos	hijo	Libro	ovarios	Impoten*	Vibrador	Homosexualidad	erección/ erecciones	impotencia
condones	opiniones	tanto	humano	Literatura	óvulos	Condon*	Juguetes	humanos	corporal / corporales	intervención
cuidado	orgasmo	otra	individuo	Macho	pareja	Preservativo*	Orgías	Impotencia	respetar/ respeto	novio
culo	Panorama	estas	infancia	Madre	pasión	Dese*	Intercambio	Info-Pene	condones	sadomasoquismo
embarazada	perfumes	sea	infertilidad	Madurez	pastillas	Infan*	Semen	informaci	Genital/ genitales	sentidos
endocrino	personas	vez	instinto	Masaje	pene	vacuna*	Contacto	interior	Masturbación	apoyo
estereotipados	placer	tienen	liberación	Masculinidad	penetración	Estereotipo*	Enfermedades	labios	Sexología	bebe
Higiene	postura	parte	lubricación	Masturbación	píldoras	Mito*	Profilaxis	Mapa	cuerpo	buen
infecciones	Preguntas	hace	masturbación	Miembro	placer	placer	Anticonceptivo	masculina	sociedad	próstata
masturbarse	productos	jóvenes	miembro	Mitos	PLAYBOY	placenter*	Organismo	Masturbaci	Testosterona	sensual
miembros	Profesorado	cuerpo	misterio	Moral	pomos	Preven*	Esperma	MICO	miedo/ miedos	amiga
ORGASMO	programa	dos	mujer	Mujer	preservativo	eyaculaci*	Vulva	morbosa	Joven	animal
pene	Recursos	momento	natural	Multidisciplinar	procrear		Útero	muslos	Embarazo/ embarazos	atrae
pervertidos	reglas	deben	necesidad	Natural	provocadores	genital	Gametos	nalgas	Excitación	cerebro
preferencia	reproductiva	niño	niveles	Naturaleza	punto G	Afectiv*	Espermatozoides	necesitan	Sensual	consultas
privado	Sexología	ni	organismo	Organismo	querer	afectiva-sexual	Fecundación	nuca	Matrimonio	hormona
problemas	Salud	humana	óvulo	Orgasmo	querida	Afecto	Coito	Obtenci	Hijos/ hijo	libertad
reproducen	Sexual	porque	pasión	Padres	relaciones sexuales	virus	Filia	obtenida	Enfermedades/ enfermedad	ultrasonido
reproductivo	sexualidad	niños	pene	Pareja	reproducción	Madre*	Fetiche	orejas	violencia	Casanova
sociedad	SIDA	castidad	penetración	Pene	reproductor	Transmisi*	Infecciones	orgasmo	Fetichista/ fetichismo	Conciencia

Apéndice D

Terminología Básica de las Sexualidades en México

TÉRMINOS

Son 125 términos por cada una de las ocho áreas. En total la *Terminología Básica de las Sexualidades en México* consta de mil entradas.

SEXUAL	472	0,29	
SEXUALIDAD	346	0,21	
SEXO	342	0,21	
HOMBRES	274	0,17	
FAMILIAR	234	0,14	
HOMBRE	211	0,13	
ANTICONCEPTIVO	210	0,13	
SEXUALES	196	0,12	
PLANIFICACIÓN FAMILIAR	177	0,11	
CUERPO	161	0,1	cuerpos(11)
ABORTO	150	0,09	
HIJOS	136	0,08	
TESTOSTERONA	127	0,08	
MÉDICO	108	0,07	médicos(59)
VARONES	108	0,07	
FECUNDIDAD	106	0,06	
MATRIMONIO	102	0,06	matrimonios(10)
EMBARAZO	91	0,06	embarazos(32)
JÓVENES	85	0,05	
FAMILIA	84	0,05	
DESEO	83	0,05	deseos(18)
MORAL	83	0,05	morales(18)
AMOR	78	0,05	amores(4)
GÉNERO	78	0,04	
PERÍODO	78	0,05	períodos(23)
PENE	75	0,05	penes(6)
MEDICINA	72	0,04	medicinas(21)
EDUCACIÓN	70	0,04	
DERECHOS REPRODUCTIVOS	70	0,04	
FEMENINA	65	0,04	
VIOLENCIA	64	0,04	
ADOLESCENTE	60	0,04	adolescentes(53)
VIOLACIÓN	57	0,03	violaciones(9)
ESTERILIZACIÓN	55	0,03	
POSIBLE	55	0,03	
MÉTODOS ANTICONCEPTIVOS	55	0,03	
DEMOGRÁFICA	54	0,03	demográficas(10)
ENFERMEDAD	54	0,03	enfermedades(35)
SIDA	54	0,03	
ACTO	50	0,03	
CIENCIA	50	0,03	ciencias(21)
HIJO	50	0,03	
SALUD REPRODUCTIVA	50	0,03	
ORGASMO	45	0,03	orgasmos(3)

RELACIONES SEXUALES	45	0,03	
FEMENINO	44	0,03	
ANTICONCEPCIÓN	42	0,03	
MÉDICA	41	0,03	
USO DE ANTICONCEPTIVOS	39	0,02	
PREPUCIO	38	0,02	
ERÓTICA	37	0,01	eróticas(11)
HIMEN	37	0,02	
HORMONAS	37	0,02	
DEMOGRAFÍA	35	0,02	
HOMBRES Y MUJERES	35	0,02	
DEMOGRÁFICO	34	0,02	demográficos(14)
VIDA SEXUAL	34	0,02	
VIRGINIDAD	33	0,02	
ABSTINENCIA	32	0,02	
ANTICONCEPTIVA	32	0,02	
HORMONAL	32	0,02	hormonales(17)
MIEMBRO	31	0,02	miembros(19)
CONYUGAL	30	0,02	conyugales(4)
TRANSMISIÓN	30	0,02	
VARÓN	30	0,02	
ANATOMÍA	29	0,02	
ANDROPAUSIA	29	0,02	
CONCEPCIÓN	29	0,02	
EROTISMO	29	0,02	
CÓNYUGE	28	0,02	cónyuges(19)
NATALIDAD	28	0,02	
PÍLDORA	28	0,02	píldoras(8)
EMBARAZADA	27	0,02	embarazadas(17)
COMPORTAMIENTO SEXUAL	26	0,01	
ESPOSO	25	0,02	esposos(4)
BESO	24	0,01	besos(13)
DOLOR	24	0,01	dolores(12)
FÉRTIL	24	0,01	fértiles(8)
CÓDIGO PENAL	24	0,01	
ERECCIÓN	23	0,01	
ESPOSA	23	0,01	esposas(6)
ESTERILIZA	23	0,01	
IMPOTENCIA	23	0,01	
SEXUALMENTE	23	0,01	
VIHSIDA	23	0,01	
ABORTAR	22	0,01	abortara(2)
ACTOS	22	0,01	
COITO	22	0,01	
DIU	22	0,01	
GLANDE	22	0,01	
HOMOSEXUAL	22	0,01	homosexuales(12)
PRÁCTICAS SEXUALES	22	0,01	
EMOCIONAL	21	0,01	emocionales(6)
GENITAL	21	0,01	genitales(14)

ADULTO	20	0,01	adultos(14)
CLÓN	20		
LÍBIDO	20		
LIBERTAD SEXUAL	20	0,01	
VIDA HUMANA	20	0,01	
PARA LAS MUJERES	20	0,01	
CIENTÍFICO	19		
CONCEBIDO	19	0,01	concebidos(6)
DESEADO	19	0,01	deseados(11)
DIVORCIO	19	0,01	
HIJA	19	0,01	hijas(11)
MENSTRUACIÓN	19	0,01	
COMPORTAMIENTO REPRODUCTIVO	19	0,01	
DERECHOS HUMANOS	19	0,01	
MEDICINAS COMPLEMENTARIAS	19	0,01	
SERES HUMANOS	19	0,01	
BIOLÓGICO	18	0,01	biológicos(7)
CÁNCER	18	0,01	
FETO	18	0,01	fetos(7)
MARIDO	18	0,01	maridos(6)
NACIMIENTOS	18	0,01	
VIRILIDAD	18	0,01	
EDUCACIÓN SEXUAL	18	0,01	
RELACIÓN SEXUAL	18		
ANORMAL	17	0,01	anormales(5)
CAMA	17	0,01	
SALUD SEXUAL	17	0,01	
TRANSMISIÓN SEXUAL	17	0,01	
ESTERILIZADAS	16		
NACIDOS	16		
PENETRACIÓN	16		
SATISFACCIÓN	16		
VIH	16		
VIOLADA	16	violadas(4)	
SEXUAL Y REPRODUCTIVA	16		
SEXUALIDAD EN MÉXICO	16		
BEBÉ	15		
CLÍNICA	15	clínicas(6)	
ERECCIONES	15		
ERÓTICO	15	eróticos(9)	
ESTERILIZAR	15	esterilizarse(2)	
FISIOLOGÍA	15		
EFECTOS SECUNDARIOS	15		
CIENTÍFICA	14	científicas(8)	
CORAZÓN	14		
DESEA	14		
ÉSTOS	14		
HORMONA	14		
INCESTO	14		
INYECCIÓN	14		

MACHO	14	machos(2)
MARITAL	14	
SEXOS	14	
ÚTERO	14	
EDAD FÉRTIL	14	
EFECTOS COLATERALES	14	
ADOLESCENCIA	13	
CICLO	13	ciclos(2)
CIRCUNCISIÓN	13	
CONDÓN	13	
DISPOSITIVO	13	dispositivos(3)
DOCTOR	13	doctora(3)
EDUCATIVO	13	educativos(4)
ESTRADIOL	13	
EYACULACIÓN	13	
GÉNEROS	13	
HETEROSEXUAL	13	heterosexuales(5)
MACHISTA	13	machistas(7)
MASTURBACIÓN	13	
ÓRGANO	13	órganos(8)
ÓVULO	13	
PARIR	13	
SATISFACER	13	satisfacerse(2)
TRANSICIÓN	13	
CUERPO FEMENINO	13	
NIVELES DE TESTOSTERONA	13	
ADOPCIÓN	12	
ADULTA	12	adultas(7)
ERÉCTIL	12	
FEMINISTA	12	feministas(2)
IMPLANTE	12	implantes(9)
INYECCIONES	12	
INYECTABLE	12	inyectables(5)
NATAL	12	
ORAL	12	orales(6)
PROSTITUTA	12	prostitutas(6)
ACTO SEXUAL	12	
CUERPO HUMANO	12	
DELITOS SEXUALES	12	
DESEO SEXUAL	12	
VIOLENCIA DOMÉSTICA	12	
ADULTERIO	11	
DIAGNÓSTICO	11	diagnósticos(6)
DISCRIMINACIÓN	11	
DISFRUTAR	11	disfrutarlo(2)
ENFERMO	11	enfermos(5)
HOMOSEXUALIDAD	11	
MEDICAMENTO	11	medicamentos(7)
MENOPAUSIA	11	
ABSTINENCIA SEXUAL	11	

CIENCIAS SOCIALES	11	
DEMOGRAFÍA HISTÓRICA	11	
ESE MOMENTO	11	
MATERNO-INFANTIL	11	
REEMPLAZO DE HORMONAS	11	
AMOROSO	10	amorosos(4)
COITAL	10	coitales(7)
DESEAN	10	
DISFRUTE	10	disfrutes(3)
DISFUNCIÓN	10	
EPIDEMIA	10	epidemias(3)
FÁRMACO	10	fármacos(5)
FEMENINOS	10	
FIGURA	10	
INFECCIÓN	10	
INSATISFECHA	10	insatisfechas(2)
MORALIDAD	10	
VAGINAL	10	vaginales(3)
CARACTERÍSTICA DEMOGRÁFICA	10	
CÓDIGO CIVIL	10	
HIJOS Y	10	
IDENTIDAD SEXUAL	10	
INTERVALOS INTERGENÉSICOS	10	
MEDICINA MEDIOAMBIENTAL	10	
MÉTODO ANTICONCEPTIVO	10	
MINISTERIO PÚBLICO	10	
SISTEMA FAMILIAR	10	
DEMANDA NO SATISFECHA	10	
TERAPIA DE REEMPLAZO	10	
CARNAL	9	
CIENTÍFICOS	9	
CIRCUNCIDADO	9	circuncidados(5)
CONTRACEPTIVO	9	contraceptivos(3)
CORPORAL	9	corporales(4)
EDUCADA	9	educadas(5)
EMBARAZADOS	9	
EMBRIÓN	9	
ERECTO	9	
FECUNDACIÓN	9	
HÍMENES	9	
INOCENTE	9	inocentes(2)
LACTANCIA	9	
MÉDICAS	9	
TESTÍCULOS	9	
VIRGEN	9	
CÓDIGOS PENALES	9	
DERECHOS SEXUALES	9	
MEDIO RURAL	9	
MUJERES JÓVENES	9	
PLACER SEXUAL	9	

PRÁCTICA ANTICONCEPTIVA	9	
PRIMERA VEZ	9	
RESPUESTA SEXUAL	9	
SEXUALIDAD MASCULINA	9	
CARICIA	8	caricias(6)
CELO	8	celos(3)
CLÍTORIS	8	
CUPIDO	8	
DESCENDENCIA	8	
DOCTORES	8	
FECUNDADO	8	
HEMBRA	8	hembras(3)
HUMOR	8	
ÍMPETU	8	
INCIRCUNCISO	8	
NACIDO	8	
AGUA POTABLE	8	
EXPERIENCIAS SEXUALES	8	
PAREJAS UNIDAS	8	
PRIMERA UNIÓN	8	
SALUD PÚBLICA	8	
SEXUALIDAD FEMENINA	8	
SEXUALIDAD HUMANA	8	
VIOLACIÓN SEXUAL	8	
CUERPO DEL PENE	8	
PARA EL HOMBRE	8	
SEXUALES Y REPRODUCTIVOS	8	
AFECTO	7	
BELLA	7	
BIOLÓGICA	7	
CASTIDAD	7	
CONCEPCIONES	7	
DOMINACIÓN	7	
DOSIS	7	
EDUCACION	7	
EMOCIONES	7	
ENFERMERA	7	enfermeras(5)
ESTIMULANTES	7	
FISIOLÓGICO	7	fisiológicos(3)
FLÁCIDO	7	
GENÉTICA	7	
MATRIMONIAL	7	matrimoniales(3)
NACER	7	
PRÓSTATA	7	
PUBERTAD	7	
SEMEN	7	
SEXOLOGÍA	7	
VACUNAS	7	
VIRTUD	7	
CONDUCTA HUMANA	7	

CONTROL NATAL	7	
DEMANDA INSATISFECHA	7	
FUNCIÓN SEXUAL	7	
MUJER VIOLADA	7	
POBLACIÓN FEMENINA	7	
PROCESO REPRODUCTIVO	7	
TRANSICIÓN DEMOGRÁFICA	7	
CONTRA DEL SEXO	7	
SEXO POR PLACER	7	
USO DEL CONDÓN	7	
ABANDONADO	6	abandonados(2)
ABANDONO	6	
AMANTE	6	amantes(2)
ANESTESIA	6	
BEBÉS	6	
CLÍNICO	6	clínicos(3)
CONASIDA	6	
CONCUBINATO	6	
CÓPULA	6	
CUELLO	6	
CURA	6	curas(2)
DEBILIDAD	6	
DESEABA	6	deseaban(3)
DESEADA	6	deseadas(3)
ESCROTO	6	
ESPERMATOZOIDE	6	
ESTERILIZADO	6	esterilizados(2)
ESTUPRO	6	
EUNUCO	6	
EXCITACIÓN	6	
FANTASÍAS	6	
FIDELIDAD	6	
FORNICACIÓN	6	
GAY	6	
GINECÓLOGO	6	ginecólogos(2)
INMORAL	6	
INTENSA	6	intensas(3)
ÍNTIMA	6	
INTRAUTERINO	6	
LABIOS	6	
MENOPÁUSICAS	6	
PARTERA	6	parteras(3)
PATOLÓGICOS	6	
PEDOFILIA	6	
PIERNAS	6	
PORNOGRAFÍA	6	
PRESERVATIVO	6	preservativos(4)
SEXUALIDAD;	6	
TRANSEXUAL	6	transexuales(2)
TRANSGRESIÓN	6	

VIOLADOR	6	violadores(2)
VIRUS	6	
ACTIVIDAD SEXUAL	6	
ATENCIÓN MÉDICA	6	
CONDUCTA SEXUAL	6	
CORREAS ACOLCHADAS	6	
DISFUNCIÓN ERÉCTIL	6	
DISPOSITIVO INTRAUTERINO	6	
ESTUDIOS DEMOGRÁFICOS	6	
PÍLDORA ANTICONCEPTIVA	6	
POBLACIÓN MASCULINA	6	
PRIMER PARTO	6	
RELACIONES COITALES	6	
SERVICIOS MÉDICOS	6	
VIOLENCIA INTRAFAMILIAR	6	
CAMA CON CORREAS	6	
CONTRA LAS MUJERES	6	
CONTROL DE NATALIDAD	6	
DEMANDA POR ANTICONCEPTIVOS	6	
EMBARAZOS NO DESEADOS	6	
PARA LA PREVENCIÓN	6	
PARTES DEL CUERPO	6	
PROGRAMA DE PLANIFICACIÓN	6	
USUARIAS DE ANTICONCEPTIVOS	6	
ABORTA	5	
ABORTIVAS	5	
ABORTIVO	5	abortivos(3)
ADOPTAR	5	
AMOROSA	5	amorosas(2)
ANDRÓGENO	5	andrógenos(2)
ANORGASMIA	5	
ANORGÁSMICAS	5	
BISEXUALIDAD	5	
CONVIVENCIA	5	
CROMOSOMAS	5	
DISCRIMINAR	5	
EDUCAR	5	
ENCINTA	5	encintas(3)
ENVEJECER	5	
ESPERMA	5	
ESTIMULAR	5	
ESTRÓGENO	5	
FETAL	5	
FIEL	5	fieles(3)
FLUIDO	5	
GAYS	5	
GLÁNDULA	5	glándulas(2)
GOZAR	5	
INCONTINENCIA	5	
INNATA	5	innatas(2)

INTIMIDAD	5	
LESBIANAS	5	
MONOGAMIA	5	
MORALISTA	5	moralistas(2)
MORALMENTE	5	
NOVIA	5	
NOVIO	5	novios(2)
PROSTÁTICA	5	
SÍNDROME	5	
TRANSGRESIONES	5	
ABUSO SEXUAL	5	
CIENCIA MÉDICA	5	
CONTROLES SOCIALES	5	
DEMOGRAFÍA MÉDICA	5	
ENFERMEDAD PROSTÁTICA	5	
ESTUDIOS CUALITATIVOS	5	
EXCITACIÓN SEXUAL	5	
HOMBRES CIRCUNCIDADOS	5	
LEGISLACIÓN PENAL	5	
MÉTODOS AGREGATIVOS	5	
MÉTODOS HORMONALES	5	
MÉTODOS MODERNOS	5	
MORAL SEXUAL	5	
PAREJA SEXUAL	5	
PREFERENCIAS REPRODUCTIVAS	5	
PRIMER HIJO	5	
PRIMERA CONCEPCIÓN	5	
RELACIONES FAMILIARES	5	
TORRENTE SANGUÍNEO	5	
VIOLENCIA SEXUAL	5	
ZONAS RURALES	5	
CONTRA DEL ABORTO	5	
CONTRA LA MUJER	5	
NIVELES DE HORMONAS	5	
PARA LA POBLACIÓN	5	
PIEL DEL PENE	5	
PRACTICAR LA ABSTINENCIA	5	
SALUD MATERNO-INFANTIL	5	
ABANDONAR	4	abandonarse(2)
ADOPTANDO	4	
AFRODITA	4	
AMADA	4	
ANTIFECUNDATIVO	4	
BELLO	4	
BIOLOGÍA	4	
BIOLOGICAS	4	
BISEXUAL	4	bisexuales(2)
CONCUBINA	4	
CONCUBINARIO	4	
CONDONES	4	

CONTAMINACIÓN	4	
CRIANZA	4	
CRIATURA	4	
DISCRIMINACION	4	
DISFUNCIONES	4	
DIVERTIRME	4	
EDUCADO	4	
ENVEJECIMIENTO	4	
EROS	4	
ESPALDA	4	
ÉSTERES	4	
ESTERILIDAD	4	
ESTIMULA	4	
ESTÍMULOS	4	
FERTILIDAD	4	
FLUJO	4	
GOZO	4	
HOMOFOBIA	4	
IMPOTENTE	4	impotentes(2)
INMUNOLÓGICA	4	inmunológicas(2)
INTIMATE	4	
LENGUA	4	lenguas(2)
MAMA	4	
MASTURBARSE	4	
MEMBRANA	4	
MONÓGAMA	4	monógamas(2)
MUCOSA	4	mucoso(2)
NUPCIALIDAD	4	
PEDÓFILO	4	pedófilos(2)
PROGEVERA	4	
PROSTITUCIÓN	4	
SENSUAL	4	
SEXUAL;	4	
TRANSGÉNERO	4	
UTERINA	4	
VACUNACIÓN	4	
VAGINA	4	vaginas(3)
ABORTO PROVOCADO	4	
ANATOMÍA GENITAL	4	
ANATOMÍA HUMANA	4	
ANTICONCEPCIÓN HORMONAL	4	
ANTICONCEPTIVOS MODERNOS	4	
ASPECTOS FUNDAMENTALES	4	
CASA PATERNA	4	
CONDICION NATURAL	4	
CONDICIONES BIOLOGICAS	4	
CONSTRUCCIÓN SOCIAL	4	
CONSTRUCCIONES SOCIALES	4	
CONTEXTO CULTURAL	4	
CONTROL DEMOGRÁFICO	4	

CONTROL SOCIAL	4
CUERPO MASCULINO	4
CULTURA SEXUAL	4
DIAGNÓSTICOS SOCIODEMOGRÁFICOS	4
DISFUNCIÓN SEXUAL	4
ENCUENTROS SEXUALES	4
ENERGIA SEXUAL	4
ESTERILIZACIÓN FEMENINA	4
FIBRAS MUSCULARES	4
HAMBRE CARNAL	4
HOMBRE-MUJER	4
IMPULSO SEXUAL	4
INSTINTO SEXUAL	4
LOCALIDADES RURALES	4
MEDIO URBANO	4
MENOPAUSIA MASCULINA	4
MORTALIDAD INFANTIL	4
MOVILIDAD SOCIAL	4
MUJERES ADOLESCENTES	4
MUJERES CASADAS	4
MUJERES ENCUESTADAS	4
NACIMIENTOS EVITADOS	4
NUEVO EMBARAZO	4
ORGANIZACIÓN FAMILIAR	4
ÓRGANOS SEXUALES	4
PENSAMIENTOS SEXUALES	4
PENSIÓN ALIMENTICIA	4
PREPUCIO INTERNO	4
PRIMERA RELACIÓN	4
PROCESOS REPRODUCTIVOS	4
PUNTO G	4
REPRESIÓN SEXUAL	4
REVOLUCIÓN SEXUAL	4
SALUD FAMILIAR	4
SISTEMA HORMONAL	4
SITUACIONES SEXUALES	4
TESTOSTERONA LIBRE	4
VIDA CONYUGAL	4
VIDAS REPRODUCTIVAS	4
ABORTO EN MÉXICO	4
ANTICONCEPTIVO EN MÉXICO	4
CABEZA DEL PENE	4
CÁNCER DE MAMA	4
CASOS DE SIDA	4
DELITOS DE INCONTINENCIA	4
DEMANDA POR PLANIFICACIÓN	4
ÉSTERES DE TESTOSTERONA	4
INFECCIÓN POR VIH	4
PACIENTES CON VIHSIDA	4
PARA EL PERÍODO	4

PARA LOS HOMBRES	4
PARA LOS INDIVIDUOS	4
PREVENCIÓN DE ENFERMEDADES	4
PRIMERA RELACIÓN SEXUAL	4
PROCREAR ES FORNICACIÓN	4
PROMEDIO DE HIJOS	4
REEMPLAZO DE TESTOSTERONA	4
RESPUESTA SEXUAL HUMANA	4
SEXUALIDAD Y GÉNERO	4
TRANSMISIÓN SEXUAL ETS	4
ABORCIÓN	3
ABORTAN	3
ABSTENCIONISMO	3
ABSTENERSE	3
ADULTEZ	3
AFECTIVA	3
AMADO	3
AMENORREA	3
AMIGABLE	3
ANÓMALOS	3
BIOLOGICA	3
BIOLÓGICAS	3
BIOLOGICISTA	3
BIÓLOGOS	3
CANCERÍGENO	3
CAPUCHÓN	3
CINTURA	3
CLONES	3
COMADRONAS	3
CONCEBIR	3
CONCIBE	3
CONCUBINOS	3
CONDON	3
CONTAGIO	3
CONTRACCIÓN	3
CORINTIOS	3
CORRELACIÓN	3
CULOS	3
DESEAS	3
DESEE	3
DESNUDO	3
DESVIACIONES	3
DIAGNÓSTICA	3
DIVORCIADO	3
DOLOROSA	3
DONCELLA	3
EMBRIONES	3
EMOCION	3
ENFERMAS	3
ENROJECIMIENTO	3

ENVEJECEN	3
EPIDEMIOLÓGICA	3
ESMEGMA	3
ESTÉRILES	3
ESTERILIZANTES	3
ESTIMULADO	3
FEMENINAS	3
FEMINIDAD	3
FETOS,	3
FRICCIÓN	3
GANAS	3
GENEALOGÍAS	3
HEMORRAGIAS	3
HERMOSO	3
HÚMEDOS	3
INFECCIONES	3
INFECTADAS	3
INFIDELIDAD	3
INSATISFECHOS	3
JERINGAS	3
JOVENCITAS	3
MORBILIDAD	3
NACE	3
ORINAR	3
PATOLÓGICA	3
PLACENTARIO	3
PORNO	3
PREMATRIMONIAL	3
PROGENITORES	3
PROMISCUIDAD	3
PÚBERES	3
SEXISTAS	3
SEXÓLOGOS	3
SEXUALIDADES	3
TÉTANOS	3
TRANSFORMATION	3
VACUNACIONES	3
VENEREAS	3
VIRGINAL	3
VOYEURISMO	3
ABORTO PARCIAL	3
ABORTOS CLANDESTINOS	3
ACTIVIDAD HUMANA	3
ACTO SIMBÓLICO	3
ANATOMÍA SEXUAL	3
ANTICONCEPCIÓN DURA	3
APOYO EMOCIONAL	3
ÁREA CULTURAL	3
ASISTENCIA MÉDICA	3
AVANCES FEMENINOS	3

CARÁCTER TEÓRICO	3
CLASE MEDIA	3
CLASE SOCIAL	3
CÓDIGO MORAL	3
COMPORTAMIENTO HUMANO	3
CONDUCTA REPRODUCTIVA	3
CONDUCTAS SEXUALES	3
CONSECUENCIA EXTREMA	3
CONTEXTOS RURALES	3
CONTRAER MATRIMONIO	3
CÓNYUGE INOCENTE	3
CRECIMIENTO DEMOGRÁFICO	3
DERECHO ABSOLUTO	3
DERECHO EXCLUSIVO	3
DESEO ANIMAL	3
ENERGIA VITAL	3
ENFERMEDADES VENEREAS	3
ERECCIONES ESPONTÁNEAS	3
ESTERILIZACION OCULTAS	3
ESTUDIOS MONOGRÁFICOS	3
EXPERIENCIA PERSONAL	3
EXPERIENCIA SEXUAL	3
EXPERIENCIAS ERÓTICAS	3
FUNCIÓN ERÉCTIL	3
HOMBRES COMO	3
HOMBRES MAYORES	3
HORMONALES ORALES	3
HOSTIGAMIENTO SEXUAL	3
IDENTIDAD FEMENINA	3
IDENTIDAD MASCULINA	3
INVESTIGACIÓN ANTROPOLÓGICA	3
JÓVENES CASADOS	3
JÓVENES MUJERES	3
MACHISTAS INSPIRADOS	3
MADRES ADOLESCENTES	3
MÉDICOS LEGISTAS	3
MEDIDAS PREVENTIVAS	3
METAS DEMOGRÁFICAS	3
MORTALIDAD MATERNA	3
MUJERES ENTREVISTADAS	3
OPCIONES ANTICONCEPTIVAS	3
OPERACIÓN FEMENINA	3
ÓRGANOS GENITALES	3
PARIR BEBÉS	3
PATRIMONIO FAMILIAR	3
PERMISIBLE ABORTAR	3
PERSONAS TRANSGÉNERO	3
PIEL SECA	3
POSTURA INICIAL	3
PRIMER COITO	3

PRIMER EMBARAZO	3
PROBLEMAS SEXUALES	3
PROCESOS DEMOGRÁFICOS	3
RECONOCIMIENTO PLACENTARIO	3
RÉGIMEN DEMOGRÁFICO	3
REGIÓN GENITAL	3
RELACIONES OCASIONALES	3
ROLES SEXUALES	3
SALUD MENTAL	3
SATISFACCIÓN SEXUAL	3
SEXO COMERCIAL	3
SEXO OPUESTO	3
SEXUALMENTE SENSIBLES	3
SUBORDINACIÓN FEMENINA	3
VIH-SIDA	3
VIOLACIONES SEXUALES	3
VIOLENCIA FÍSICA	3
ZONAS URBANAS	3
ÁMBITOS DE FAMILIA	3
ANATOMÍA Y FISIOLOGÍA	3
ANTICONCEPTIVA DE EMERGENCIA	3
ANTICONCEPTIVO Y PÍLDORAS	3
BRUTAL DESEO ANIMAL	3
CÁNCER DE PRÓSTATA	3
CÁNCER DEL CUELLO	3
CAUSAS DE MUERTE	3
CONSUMO DE ESTIMULANTES	3
CONTRA EL PUDOR	3
CONTROL DE POBLACIÓN	3
CUELLO DEL ÚTERO	3
DESPUÉS DEL MATRIMONIO	3
DESPUÉS DEL PARTO	3
DESPUÉS DEL PROGRAMA	3
DOLORES Y ACHAQUES	3
EDAD DE FORMACIÓN	3
ESTUDIO DEL HIMEN	3
FORMAS DEL HIMEN	3
GRUPO DE MUJERES	3
GRUPOS DE MUJERES	3
HISTORIA DEL CUERPO	3
INFORMACIÓN Y EDUCACIÓN	3
INSTRUMENTO DE PLACER	3
NIVEL DE TESTOSTERONA	3
NIVELES DE ESTRADIOL	3
NORMA SOBRE VIHSIDA	3
PARA LA PLANEACIÓN	3
PARA LA PROCREACIÓN	3
PELIGRO DE VIOLACIÓN	3
PENALIZACIÓN DEL ABORTO	3
POLÍTICAS DE POBLACIÓN	3

POLÍTICAS DE SALUD	3
PRÁCTICAS DE RIESGO	3
PRIMERA UNIÓN MARITAL	3
PROCREAR ES PECADO	3
PRODUCCIÓN DE HORMONAS	3
PRODUCCIÓN DE TESTOSTERONA	3
PROGRAMAS DE PLANIFICACIÓN	3
PROGRAMAS DE POBLACIÓN	3
PROHÍBEN EL ABORTO	3
RELACIONES ENTRE HOMBRES	3
RELACIONES HOMBRE-MUJER	3
RESULTADO DEL INCESTO	3
SEXO Y SEXUALIDAD	3
SIDA Y ENFERMEDADES	3
SISTEMA DE SALUD	3
SOBREVIVE SIN SEXO	3
TRABAJADORAS DEL SEXO	3
VIOLACIÓN ENTRE CÓNYUGES	3
ABANDONADAS	2
ABDOMEN	2
ABORTADOS,	2
ABSTENCIÓN	2
ABSTENERTE	2
ACOPLAMIENTO	2
ADAN	2
ADOLESCENTES;	2
AFECTIVIDAD	2
ANAL	2
ANATÓMICAS	2
ANATOMISTA	2
ANDRÓGENA	2
ANO	2
ANTICUERPOS	2
ARDIENTES	2
ARDOR	2
AUTOERÓTICA	2
BELLEZA	2
BÍOLÓGICOS	2
CELIBATO	2
CONCUBINAS	2
CONSORTE	2
CONTAGIEN	2
CONTRACEPCIÓN	2
CÓNYUGUE	2
CUNA	2
DAMAS	2
DEPRAVADAS	2
DESEABLES	2
DESFLORACIÓN	2
DESFOGUE	2

DESNUDAS	2
EMBARAZA	2
EMBARAZAN	2
ENAMORARON	2
ENGENDRARA	2
EPIDEMIOLOGÍA	2
ERÓGENAS	2
ESPERMATICIDAS	2
ESTERILES	2
ESTIRPE	2
EXCITA	2
EXCITABILIDAD	2
EXCITANTE	2
EXHIBICIONISMO	2
EXPLORAR	2
EXPLORATORIO	2
EXPLORE	2
FALOPIO	2
FETICHISMO	2
FISIOLÓGICA	2
FISIOLOGICO	2
GENÉTICO	2
GINECO	2
GLÚTEOS	2
GOZA	2
GOZADO	2
GOZAN	2
HIMENÓLOGOS	2
HOMBRÍA	2
IMPLANTACIÓN	2
INEXPERIENCIA	2
INFERTILIDAD	2
INHIBICIONES	2
INMUNOLÓGICO	2
INSATISFACCIÓN	2
INTIMO	2
ÍNTIMO	2
INYECTAR	2
JOVENCITOS	2
JUGUETE	2
LAPAROSCOPIA	2
LEGRADO	2
LÉSBICOS	2
LUJURIA	2
MAMARIOS	2
MASOQUISMO	2
MATRIZ	2
MEDICAL	2
MENSTRUAL	2
MESTEROLONA	2

METABOLISMO	2	
MIERDA	2	
MONÓGAMO	2	
NACIDAS	2	
NAZCA	2	
NUPCIALES	2	
OBSCENAS	2	
OBSTETRICIA	2	
OMBLIGO	2	
ORINA	2	
PAPANICOLAOU	2	papanicolau (3)
PARAFILIAS	2	
PATOLOGÍA	2	
PEDOS	2	
POLLA	2	
POSTPARTO	2	
PREMARITAL	2	
PREMENSTRUAL	2	
PRENATAL	2	
PROGENIE	2	
PROGESTERONA	2	
PROMISCUAS	2	
PROTOGENÉSICOS	2	
PUERPERIO	2	
SADISMOMASOQUISMO	2	
SALPINGOCLASIA	2	
SEXUADA	2	
SEXUADO	2	
SEXUALIS;	2	
SUBDÉRMICO	2	
TANTRA	2	
TORNEADAS	2	
TRANSEXUALISMO	2	
TRANSMISIBLES	2	
TRANSPIRACIÓN	2	
TRASVESTISMO	2	
UMBILICAL	2	
ABORTOS ILEGALES	2	
ABORTOS INDUCIDOS	2	
ABORTOS SEGUROS	2	
ACCESO SEXUAL	2	
ACCIÓN HUMANA	2	
ACCIÓN SOCIAL	2	
ACTIVIDAD COITAL	2	
ACTIVIDADES SOCIALES	2	
ACTIVO SEXUALMENTE	2	
ACTO BESTIAL	2	
ACTO IMPUESTO	2	
ADOLESCENTES PÚBERES	2	
ADULTO CHIQUITO	2	

AGENCIA CATÓLICA	2
AGENCIA INTERNACIONAL	2
AGENTES ENCUESTADOS	2
AGREGANDO ABORTIVOS	2
AMORES LÉSBICOS	2
ANATOMÍA BÁSICA	2
ANATOMÍA COMPARADA	2
ANATOMÍA MASCULINA	2
ANTI-ABORCIÓN	2
ANTICONCEPCIÓN QUIRÚRGICA	2
ANTICONCEPTIVOS HORMONALES	2
APARATO CIENTIFICO	2
ARBITRAJE MÉDICO	2
ASFIXIA AUTOERÓTICA	2
ASOCIACIÓN MÉDICA	2
ASOCIACIÓN MEXICANA	2
ASOCIACIÓN NORTEAMERICANA	2
ASPECTO MACHISTA	2
ASPECTOS ESPECÍFICOS	2
ATENCIÓN PRIMARIA	2
AUTORIDADES SANITARIAS	2
BAJA ESCOLARIDAD	2
BAJO AMENAZA	2
BAJO ANESTESIA	2
BIENESTAR FAMILIAR	2
CAMBIOS PSICOLÓGICOS	2
CAMINO CORRECTO	2
CAPACIDAD ERÓTICA	2
CAPACIDAD REPRODUCTIVA	2
CAPACIDAD SEXUAL	2
CARÁCTER RITUAL	2
CARENCIAS SOCIALES	2
CASOS NOTIFICADOS	2
CASOS PRESENTADOS	2
CASOS PROMEDIO	2
CASTIGO CELESTIAL	2
CATEGORÍAS SEXUALES	2
CENTROS MÉDICOS	2
CICLO MENSTRUAL	2
CLÍNICA OFICIAL	2
COITAL HETEROSEXUAL	2
COLECCIÓN MICROFILMADA	2
COMERCIO CARNAL	2
COMERCIO SEXUAL	2
COMPORTAMIENTO DEMOGRÁFICO	2
CONCIENCIA TRANQUILA	2
CONCURSOS MIXTOS	2
CONDICION BIOLOGICA	2
CONOTACIÓN SEXUAL	2
CONSCIENCIA PRIMIGENIA	2

CONSECUENCIAS FAVORABLES	2
CONSTRUCCIONES CULTURALES	2
CONSULTA MÉDICA	2
CONTACTO SEXUAL	2
CONTAMINACIÓN AMBIENTAL	2
CONTENIDOS CULTURALES	2
CONTEXTO EDUCATIVO	2
CONTEXTO SOCIOHISTÓRICO	2
CONTRATO SOCIAL	2
CONTROL PERSONAL	2
CONTROLES COMUNITARIOS	2
CONYUGAL PROCREATIVA	2
CÓNYUGE CULPABLE	2
COSTUMBRES HUMANAS	2
CRECIMIENTO NATURAL	2
CRIMENES SEXUALES	2
DEBATE FEMINISTA	2
DENSIDAD ÓSEA	2
DERECHO NATURAL	2
DERECHOS CIVILES	2
DERECHOS LABORALES	2
DESARROLLO SEXUAL	2
DESEO ERÓTICO	2
DISFUNCIONES SEXUALES	2
DIVORCIO SOLICITADO	2
DOMINACIÓN MASCULINA	2
EDAD REPRODUCTIVA	2
EDUCACIÓN CONSERVADORA	2
ENCUENTROS AMOROSOS	2
ENCUESTAS DEMOGRÁFICAS	2
ENFERMEDAD PSICOLÓGICA	2
ENFERMEDADES SEXUALMENTE	2
ENFOQUE INTEGRAL	2
ENTORNO SOCIAL	2
ERECCIONES MATINALES	2
ERECCIONES NOCTURNAS	2
ERECTO INCIRCUNCISO	2
ESPECIALIDADES FARMACÉUTICAS	2
ESTERILIZACIÓN MASCULINA	2
ESTERILIZACIÓN MASIVA	2
ESTERILIZACION QUÍMICAS	2
ESTERILIZACIÓN QUIRÚRGICA	2
ESTUDIO SISTEMÁTICO	2
ESTUDIOS SOCIALES	2
ESTUDIOS SOCIOLÓGICOS	2
ÉTICA SEXUAL	2
EVALUACIÓN ÉTICA	2
EXPERIENCIA MEXICANA	2
EXPLOSIÓN DEMOGRÁFICA	2
EXPRESIÓN SEXUAL	2

EYACULACIÓN PREMATURATARDÍA	2
FEMINISTA ACTIVA	2
FETOS INOCENTES	2
FISIOLOGÍA HORMONAL	2
FUNCIONES INMUNOLÓGICAS	2
GINECO-OBSTETRICIA	2