



UNIVERSIDAD NACIONAL
AUTÓNOMA DE
MÉXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

**“CREACIÓN SEMI-AUTOMÁTICA DE LA BASE DE DATOS Y MEJORA DEL
MOTOR DE BÚSQUEDA DE UN DICCIONARIO ONOMASIOLÓGICO”**

T E S I S

QUE PARA OBTENER EL GRADO DE:

**MAESTRA EN CIENCIAS
(COMPUTACIÓN)**

P R E S E N T A:

LAURA ELENA HERNÁNDEZ DOMÍNGUEZ

**DIRECTOR DE TESIS: GERARDO EUGENIO SIERRA MARTÍNEZ
CODIRECTORA DE TESIS: SYLVIE RATTÉ**

México, D.F.

2012



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

De todos los instrumentos del hombre, el más asombroso es, sin duda, el libro. Los demás son extensiones de su cuerpo. El microscopio, el telescopio, son extensiones de su vista; el teléfono es extensión de la voz; luego tenemos el arado y la espada, extensiones del brazo. Pero el libro es otra cosa: el libro es una extensión de la memoria y la imaginación.

Jorge Luis Borges

Agradecimientos

Quiero agradecer a todas las personas que estuvieron —de alguna manera— conmigo durante el desarrollo de esta tesis y, si omito a alguna persona, sepa ésta que no es por falta de gratitud, sino por culpa de mi irremediable mala memoria:

A mis padres y mi hermano por toda su ayuda, comprensión y paciencia durante el desarrollo de esta tesis.

A Edgar García Cano, por su apoyo en varios de los aspectos técnicos surgidos durante el desarrollo de este trabajo, pero también por su grata compañía y apoyo en los momentos difíciles.

Al Dr. Gerardo Sierra, mi director de tesis, por su paciencia y ayuda durante la realización de la misma y por apartar siempre un tiempo en su apretada agenda para escucharme y darme nuevas ideas.

A la Dra. Sylvie Ratté, mi codirectora de tesis, por escucharme, animarme y orientarme durante mi estancia en Montreal.

A mis sinodales, la Dra. Cristina Verde, el Dr. Alfonso Medina y el Dr. Gricory Sidorov por sus correcciones y orientación.

A las personas que colaboraron con la recopilación de datos para esta tesis: mis padres y mi hermano, además de José Velasco, Paola Butrón, Javier Flores, Arlen Pérez, Malinalli Cortés, Alejandro Murrieta, Yanet Machorro y Tamara Sierra.

Finalmente, a todos los integrantes del Grupo de Ingeniería Lingüística (GIL), por su apoyo desinteresado y su compañerismo.

Este trabajo fue realizado gracias a los apoyos recibidos por parte del Consejo Nacional de Ciencia y Tecnología (CONACYT), con la beca de posgrado nacional número 231179 y del proyecto “Vocabulario básico científico de México: una investigación de sus características, componentes y difusión” con número de registro 58923.

Índice

INTRODUCCIÓN	6
CAPÍTULO 1. DICCIONARIOS DE BÚSQUEDA ONOMASIOLOGICA	8
1.1. Tesoros	8
1.2. Diccionarios de Sinónimos.....	12
1.3. Diccionarios Visuales.....	13
1.4. Diccionarios Inversos.....	15
CAPÍTULO 2. DICCIONARIO ELECTRÓNICO DE BÚSQUEDA ONOMASIOLOGICA, DEBO	19
2.1. Arquitectura general	19
2.2. Algoritmo de búsqueda.....	30
CAPÍTULO 3. PROPUESTA DEL MOTOR DE BÚSQUEDA DE DICCIONARIOS ONOMASIOLOGICOS	32
3.1. Cambios propuestos al algoritmo de búsqueda.....	32
3.2. Cambios propuestos a la base de datos.....	38
CAPÍTULO 4. GENERACIÓN SEMI-AUTOMÁTICA DE LA BASE DE DATOS DE DICCIONARIOS ONOMASIOLOGICOS.....	43
4.1. Entradas	43
4.2. Tratamiento automático de las entradas.....	49
4.3. Ingreso de las entradas a la base de datos	52
CAPÍTULO 5. DICCIONARIO ONOMASIOLOGICO DE LAS SEXUALIDADES EN MÉXICO (DOSEX-MX)	55
5.1. La base de conocimientos	55
5.2. Diccionario resultante	57
CAPÍTULO 6. EVALUACIÓN Y RESULTADOS.....	61
6.1. Criterios de evaluación.....	61
6.2. Definiciones de los usuarios	62
6.3. Evaluación	63
CONCLUSIONES	72
BIBLIOGRAFÍA.....	73
APÉNDICE A. TÉRMINOS DEL DOSEX-MX.....	76
APÉNDICE B. LISTA DE PARO DEL DOSEX-MX.....	79
APÉNDICE C. TABLAS DE EXPERIMENTACIÓN Y EVALUACIÓN DEL DOSEX-MX.....	80

Introducción

A todos nos ha pasado alguna vez que al momento de hablar o de escribir olvidamos el nombre de una palabra. Éste es un problema que se conoce comúnmente como “tener la palabra en la punta de la lengua”. Si tenemos suerte y hay alguna persona con nosotros, generalmente somos capaces de darle a otra persona nuestra propia descripción de la palabra que no recordamos pero, ¿qué sucede cuando no hay nadie más a quién preguntar?

A los diccionarios que ayudan a sus usuarios a ir de la definición al nombre del término se les conoce como diccionarios onomasiológicos. Estos diccionarios existen en varios formatos y poseen diversas maneras de abordar el problema de la palabra en la punta de la lengua.

El objetivo de la presente tesis consiste en presentar una metodología para la generación semi-automática de diccionarios onomasiológicos electrónicos especializados que funcionan con consultas escritas por el usuario en lenguaje natural, buscando además mejorar los resultados obtenidos actualmente con el método de búsqueda onomasiológica desarrollado anteriormente en el Grupo de Ingeniería Lingüística de la UNAM por el Dr. Gerardo Sierra durante en desarrollo de su tesis doctoral en la universidad de Manchester.

Las evaluaciones de la metodología aquí propuesta y de la mejora en el motor de búsqueda del diccionario onomasiológico se realizaron mediante la creación de dos diccionarios especializados en el área de sexualidad con léxico mexicano. Ambos diccionarios poseen la misma base de conocimientos, pero sus motores de búsqueda corresponden uno a la versión original del diccionario y el otro a la mejora propuesta en este trabajo. Por lo anterior, el dominio de aplicación de esta tesis pertenece al área de sexualidad, razón por la que los ejemplos se orientarán en la mayoría a esta área.

Lo sistemas desarrollados para esta tesis son listados a continuación.

- Sistema para la selección y depuración de los resultados de Describe®
- Sistema para la generación automática de diccionarios onomasiológicos (SGADO)
- Diccionario onomasiológico de sexualidad con motor de búsqueda original
- Diccionario onomasiológico de sexualidad con motor de búsqueda mejorado

En el primer capítulo de esta tesis se habla un poco de los diccionarios onomasiológicos: qué son, para qué sirven, qué tipos de diccionarios onomasiológicos electrónicos existen, cuáles son los más representativos y sus ventajas y desventajas. En el segundo capítulo se trata a fondo la propuesta del Diccionario Electrónico de Búsqueda Onomasiológica que actualmente se encuentra en el Grupo de Ingeniería Lingüística: cuál es su arquitectura general, la base de conocimientos que requiere, así como el funcionamiento de su motor de búsqueda. El tercer capítulo consiste en la presentación de una propuesta de mejora al algoritmo de búsqueda expuesto en el segundo capítulo. El cuarto capítulo presenta una metodología completa para la creación semi-automática de diccionarios onomasiológicos electrónicos partiendo de cero: desde la obtención de la base de conocimientos, el tratamiento automático de la misma y su ingreso en la base de datos especializada. El quinto capítulo muestra la información del Diccionario Onomasiológico de las Sexualidades en México, que fue creado siguiendo la misma metodología planteada en el capítulo cuatro y cuyo motor de búsqueda corresponde a la propuesta de mejora presentada en el capítulo tres. Finalmente, el capítulo 6 proporciona una evaluación de los resultados obtenidos con el nuevo motor de búsqueda propuesto contrastados con los resultados del motor de búsqueda del diccionario original.

Las principales aportaciones de esta tesis consisten en la mejora a los resultados obtenidos con el método de búsqueda onomasiológica desarrollado anteriormente en el Grupo de Ingeniería Lingüística de la UNAM y la semi-automatización de la creación de dichos diccionarios. La mejora de los resultados consiste en una reducción significativa del listado de términos resultante, evitando que la palabra objetivo sea descartada del listado de resultados y reduciendo para el usuario el tiempo de búsqueda de la palabra objetivo.

Capítulo 1. Diccionarios de búsqueda onomasiológica

De acuerdo con Baldinger (1970) hay dos tipos de diccionarios desde el punto de vista semántico: semasiológicos y onomasiológicos. El primer tipo es el que se utiliza en forma más habitual. En este tipo de diccionarios se conoce el término y se desea encontrar su significado. El segundo tipo de diccionario es una antítesis del primero, pues se utiliza cuando se comprende el significado pero se desconoce el término con el que se le denomina. Estos diccionarios intentan resolver el problema de “la palabra en la punta de la lengua” y su principal utilidad consiste en permitir que sus usuarios sean capaces de expresarse efectivamente al escribir o hablar, al proporcionarles los términos más adecuados a sus ideas.

Los diccionarios onomasiológicos pueden encontrarse en forma impresa o electrónica. Con los diccionarios impresos se dificulta la localización de la información, pues los usuarios tienen que buscar entre las páginas de estos volúmenes, guiándose —dependiendo del tipo de diccionario que se trate— por orden alfabético o por índices temáticos. Las limitaciones propias que se presentan en los materiales impresos hacen impráctica la continuación de este modelo, por lo que los diccionarios onomasiológicos electrónicos representan una opción más funcional.

Para el desarrollo de esta tesis se realizó una investigación acerca de los tipos más comunes de diccionarios onomasiológicos que se encuentran disponibles en formato electrónico y se encontró que Sierra (2000) considera cuatro tipos de diccionarios onomasiológicos: tesauros, diccionarios visuales, diccionarios de sinónimos y diccionarios inversos. A continuación se describen los más representativos clasificados de acuerdo con la forma de uso.

1.1. Tesauros

Según Shcherba (1995), los tesauros son los diccionarios onomasiológicos más antiguos que se conocen, siendo el más antiguo el Tesoro de la Lengua Castellana o Española, de Sebastián Covarrubias escrito a principios del siglo XVII (Bouzy, 1992). En ellos se ordenan por temas o ideas los términos contenidos, con lo que, para localizar una palabra mediante un tesoro, los

usuarios deben buscar desde un área muy general donde suponen que el término podría encontrarse y, a partir de ahí, moverse hasta temas más específicos, como en las ramas de un árbol, hasta encontrar el término buscado.

Un tesoro electrónico ampliamente conocido es WordNet¹ (Miller, 1995 y Fellbaum, 1998), el cual fue desarrollado en la Universidad de Princeton para la lengua inglesa y contiene una organización del léxico inglés que contempla relaciones entre vocablos no sólo de hiperonimia e hiponimia (relaciones “*es un*”, por ejemplo, *medio_de_transporte-automóvil*), sino también de sinonimia (palabras que tienen significados similares, como *automóvil-coche*), holonimia y meronimia (relaciones “*forma parte de*” y “*tiene un*”, por ejemplo, *automóvil-llanta*). Se encuentra disponible para su consulta en línea (<http://wordnet.princeton.edu>) y también cuenta con Periscope, un sistema desarrollado para su visualización fuera de línea. Este tesoro se ha adaptado para el español y otros idiomas en lo que se conoce como EuroWordNet (Vossen, 1998).

Un ejemplo de búsqueda en EuroWordNet², obtenida desde Periscope³ se muestra a continuación con la consulta de la palabra *clitoris*. Para esta consulta existen varias formas de llegar a la palabra buscada, aunque algunas requieren un mayor esfuerzo computacional. A continuación se muestran dos posibles rutas:

1. Comenzando con el término *parte del cuerpo*, se encuentra una lista de cerca de cuarenta opciones posibles. De esta lista se puede elegir el término *órgano* y a continuación, de entre 20 términos listados, puede seleccionarse *órgano eréctil*, en el que *clitoris* es fácilmente identificable, pues sólo se encuentra acompañado de *pene*.
2. La segunda manera de buscar el término *clitoris* comienza también con la consulta de *parte del cuerpo* y la selección de *órgano*, aunque de ahí se decide tomar *órganos sexuales*, para después seleccionar *genitales* y a continuación, *genitales femeninos*, donde se encuentra listado, entre otros términos, el término *vulva* entre cuyos componentes se encuentra, finalmente, *clitoris*.

¹ WordNet 1.5 – 1996-98, Princeton University.

² EuroWordNet – 1997-1998, The EuroWordNet Consortium.

³ Periscope 1.3.2 – 1998-99, Lernout&Houspie. FLAIM database engine.

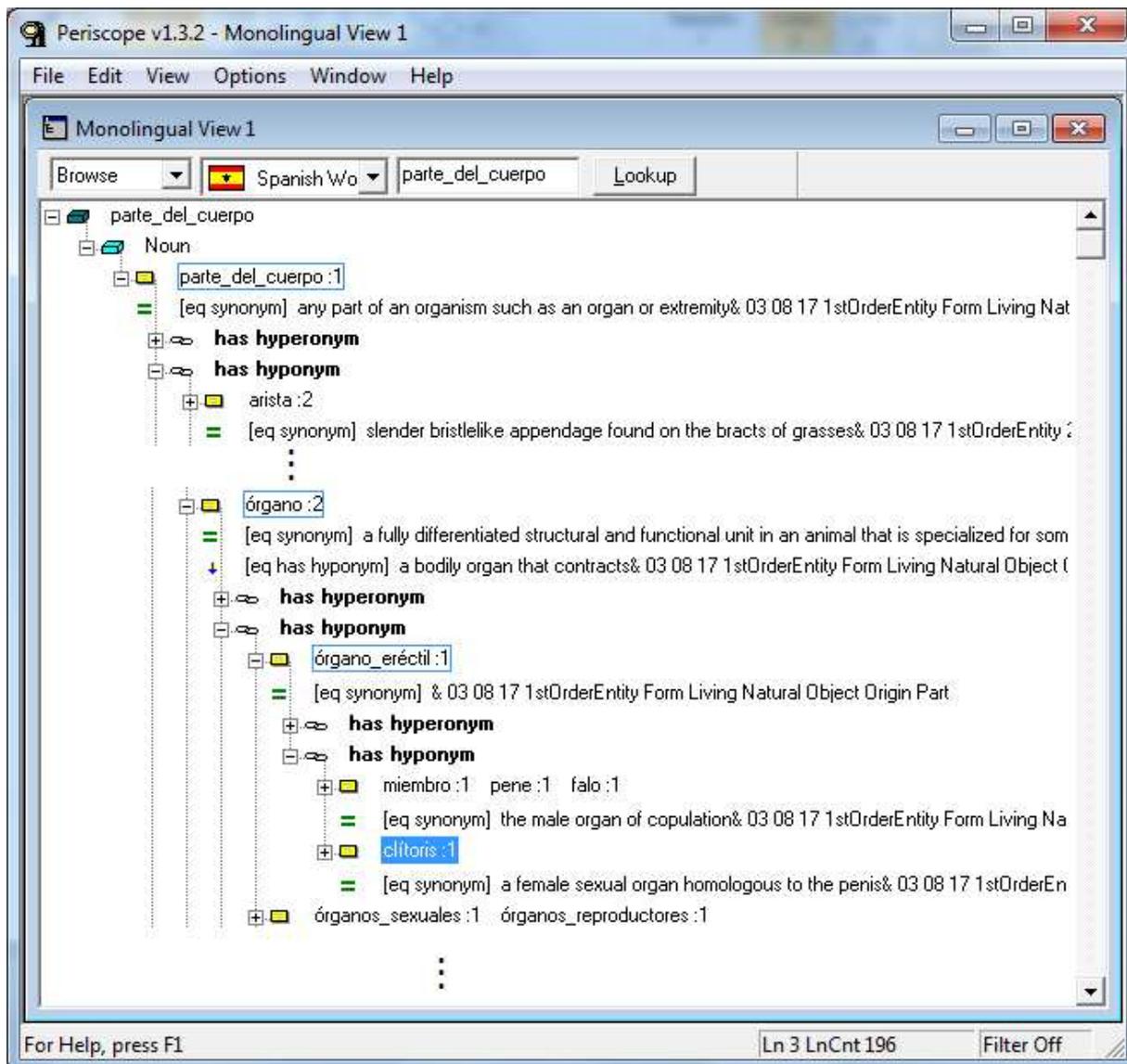


Figura 1: Esquema del primer ejemplo de búsqueda de la palabra *clitoris* en EuroWordNet

Observando estos procesos de búsqueda es posible darse cuenta de que se trata de un procedimiento secuencial de búsqueda que requiere de cierto conocimiento de la categorización específica del término que se está buscando. Por ejemplo, si por desconocimiento un usuario no considera al clitoris como un órgano eréctil, éste tardará más en encontrar la palabra que busca. También existe el problema de la falta de consenso por parte de los especialistas al momento de organizar en clasificaciones jerárquicas a una palabra, pues las clasificaciones del conocimiento no pueden traspasar las coordenadas de espacio-tiempo, dado que ellas reflejan los conocimientos y valores específicos de cada cultura y cada época (San Segundo, 1996).

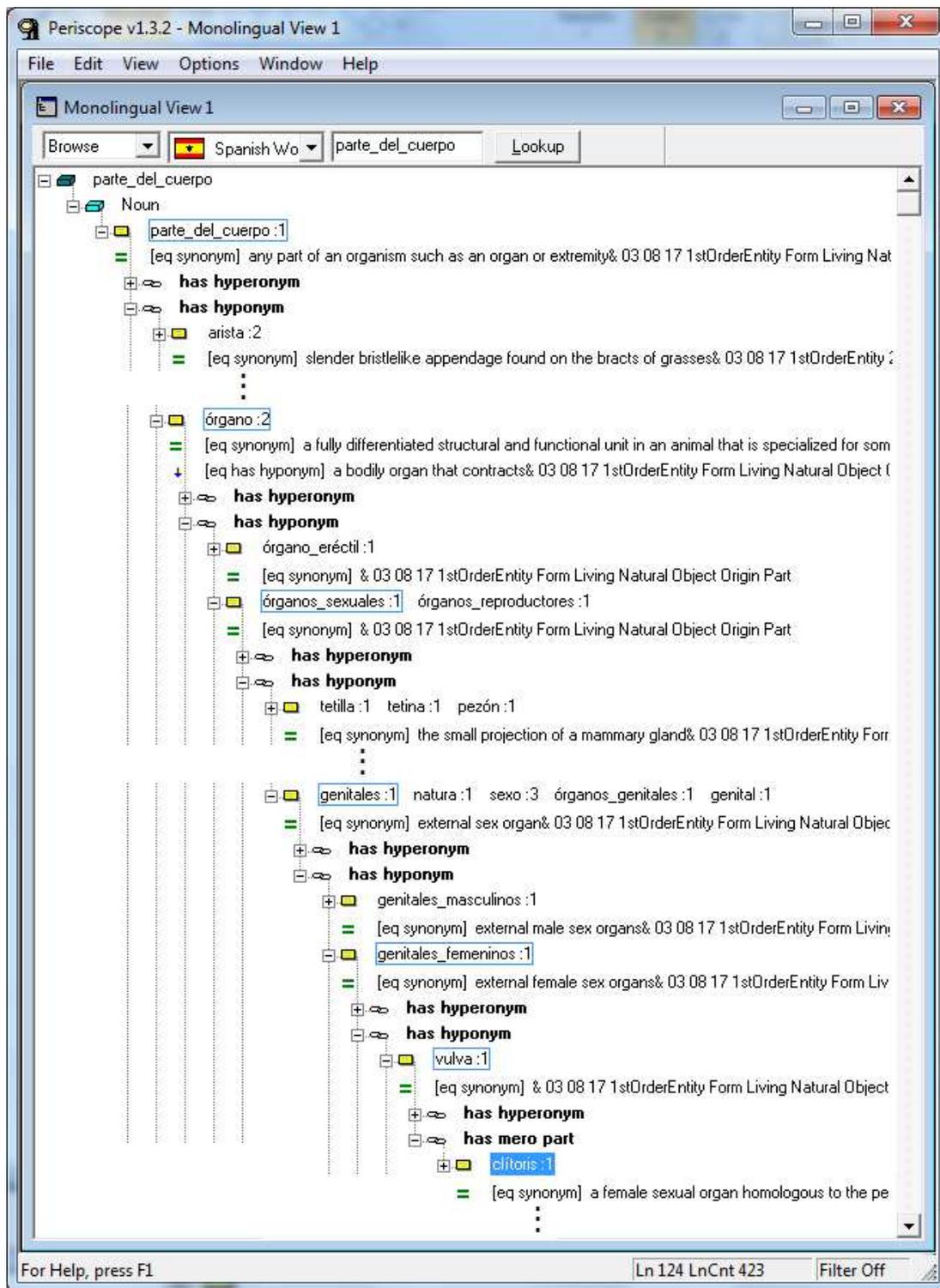


Figura 2. Esquema del segundo ejemplo de búsqueda de la palabra *clitoris* en EuroWordNet

1.2. Diccionarios de Sinónimos

Diversos autores, como Malkiel (1975), Svensén (1993) y Shcherba (1995) reconocen a los diccionarios de sinónimos como diccionarios onomasiológicos, pues permiten a los usuarios buscar un término mediante los sinónimos que recuerda en aquel momento y que tengan alguna relación semántica con el término buscado. Sin embargo, es muy posible que el usuario no recuerde tampoco algún sinónimo o que el término que busque no tenga un sinónimo.

Los diccionarios de sinónimos son listas de palabras relacionadas (semánticamente), cuyas entradas se encuentran ordenadas alfabéticamente, mientras que las palabras sinónimas pueden encontrarse ordenadas tanto alfabéticamente como con algún otro criterio (por ejemplo, por sus acepciones).

Cuando se busca una palabra en un diccionario de sinónimos es común ir saltando de un sinónimo a otro hasta encontrar el buscado. Por ejemplo, al buscar en el diccionario de sinónimos del español de WordReference⁴ la palabra *afeminado*, una posible primera palabra para comenzar la búsqueda podría ser *homosexual*, de donde obtengo los sinónimos *marica*, *maricón*, *sarasa* y *sodomita*. A continuación busco los sinónimos de *marica*, encontrando en esta entrada *homosexual* y *afeminado*.

En este ejemplo se puede observar que no siempre es posible identificar en el primer intento el grupo de sinónimos al que puede pertenecer una palabra. También existe una limitante en cuanto al vocabulario conocido por el usuario, ya que los diccionarios de sinónimos suelen contener palabras que no son muy utilizadas por la mayoría de las personas o en diferentes estratos sociales o culturales, como en el ejemplo anterior, la palabra *sarasa* y, por lo mismo, estos vocablos difícilmente serán una buena guía para que el usuario busque la palabra que olvidó.

⁴ www.wordreference.com (Michael Kellogg).

1.3. Diccionarios Visuales

Los diccionarios visuales tienen una organización temática inicial y, a partir de ahí, poseen en cada tema una serie de imágenes relacionadas con los términos que pertenecen a dicho tema o más subtemas dentro de los cuales se puede buscar.

Para ilustrar la forma de utilización de un diccionario visual, a continuación se muestra el proceso que se lleva a cabo para buscar la palabra *pezón* dentro del diccionario visual de ikonet⁵. En éste, primero se tiene una serie de temas principales, dentro de los que se elige *ser humano*, a continuación se elige el subtema *cuerpo humano*, después se puede elegir tanto *hombre>>vista anterior* como *mujer>>vista anterior*, finalmente se muestra una imagen con un hombre desnudo, visto de frente, donde se puede localizar el pezón señalado.

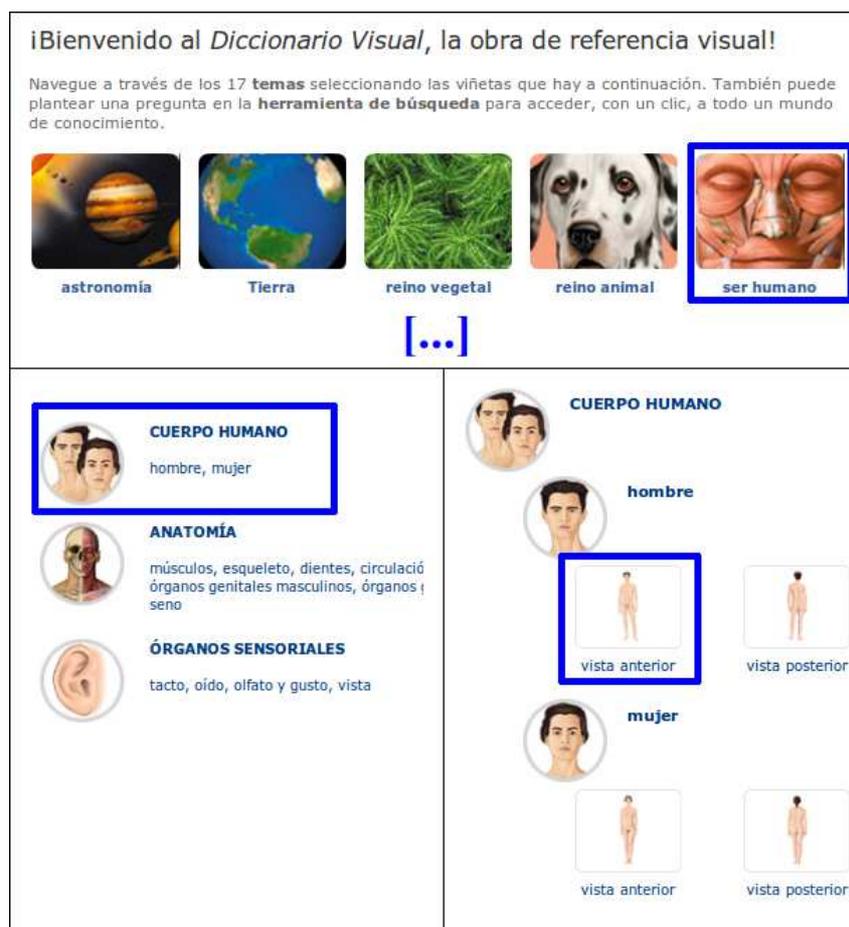


Figura 3: Búsqueda temática dentro del diccionario visual de ikonet

⁵ <http://www.ikonet.com/es/diccionariovisual/> (QA International: Diccionario Visual).

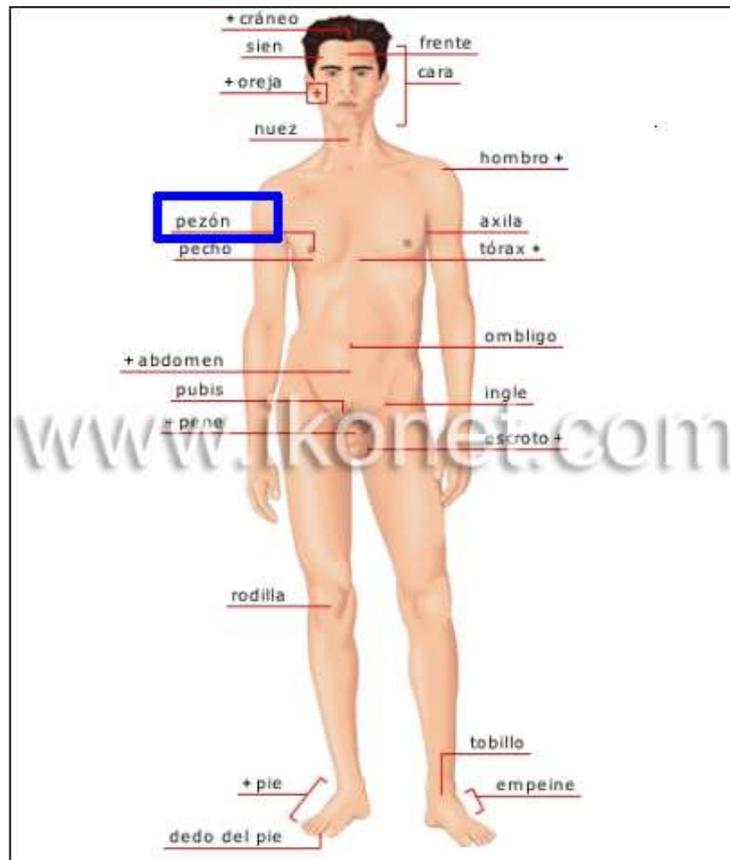


Figura 4: Localización de la palabra pezón dentro de una imagen de un diccionario visual

Como se puede apreciar, en los diccionarios visuales los usuarios localizan el término que buscan con base en la observación de una imagen que lo representa. Pueden ser muy útiles cuando la palabra corresponde a un objeto físico del mundo, pero aún así pueden existir inconvenientes pues no todas las palabras tienen una representación gráfica definida. Por ejemplo, es difícil imaginar una imagen que represente a la palabra sífilis; se trata de una palabra que es más fácil de expresar definiendo su sintomatología, o expresándola como una enfermedad. También está presente la dificultad en la organización temática de las imágenes, similar a la de los tesauros, en la que no existe un consenso, así como el hecho de que una imagen puede resultar confusa en palabras que en la mente de las personas tienen formas similares.

1.4. Diccionarios Inversos

Existen diferentes variantes de diccionarios inversos; para esta tesis se hará referencia únicamente a los diccionarios inversos onomasiológicos. Los diccionarios inversos permiten a los usuarios realizar consultas a partir de la definición del término que buscan.

1.4.1. Diccionarios de búsquedas mediante expresiones regulares

Estos diccionarios permiten al usuario ingresar palabras que él considere que podrían encontrarse en la definición de un término dentro de un diccionario, permitiéndole al usuario utilizar expresiones regulares para enlazar estas palabras, de forma que se tiene una búsqueda más sofisticada e interactiva que con los diccionarios anteriores.

Un ejemplo de búsqueda con este tipo de diccionarios se muestra a continuación con el diccionario VOX⁶ en su versión electrónica para el Diccionario General de la Lengua Española y el Diccionario Avanzado de Sinónimos y Antónimos de la Lengua Española. Al tratar de buscar la palabra *sífilis* en su módulo de búsqueda múltiple, se ingresó la consulta:

ENFERMEDAD .Y SEXUAL

Donde *.Y* corresponde al símbolo representativo de la operación de conjunción, con lo que en esta búsqueda se desean aquellos términos que contengan tanto la palabra *enfermedad* como *sexual* en su definición. No obstante, de esta consulta únicamente se obtienen los resultados: *enfermedad* y *sida*. A continuación se vuelve a hacer la prueba, esta vez con la entrada:

ENFERMEDAD .Y VENEREA

Para esta entrada no se obtuvieron resultados. Se toma entonces la forma ortográfica correcta de la palabra *venérea*, siendo la entrada final:

ENFERMEDAD .Y VENÉREA

Para la que, finalmente, se obtienen las palabras *enfermedad* y *sífilis*.

Con este ejemplo se evidencian diversos problemas, como que su uso no es intuitivo para los usuarios que desconocen la utilización de las expresiones regulares. También existe el problema

⁶ <http://www.vox.es/> (Grupo Anaya)

de que la mayoría de estos diccionarios pretende que el usuario ingrese las palabras exactas, con la ortografía correcta con las que el diccionario posee las definiciones, por lo que es difícil que usuarios no especialistas u ocasionales encuentren rápidamente el término que buscan.

1.4.2 Diccionarios de búsquedas en lenguaje natural

Para este tipo de diccionarios, el usuario introduce su propia definición en lenguaje natural, es decir, en el lenguaje con el que habla normalmente con cualquier otra persona y el sistema busca aquellos términos que coinciden con la definición que el usuario ha proporcionado. Éstos son los diccionarios cuya utilización es más fácil para el usuario, pues él sólo tiene que expresar su propio concepto de la forma en que se lo expresaría a alguien más. Pocos trabajos de este tipo se han realizado para el idioma español.

Para el idioma Japonés se tiene el trabajo de Bilac *et al.* (2004) en esta área, cuyo diccionario permite a los usuarios ingresar sus definiciones en forma libre. Este diccionario trabaja con una medida de similitud entre los conceptos basándose en comparaciones de palabras. Los autores reconocen que un problema en este diccionario es que si los usuarios ingresan en su definición palabras diferentes a aquellas con las que cuenta su base de conocimientos, aún siendo sinónimas, la medida de similitud disminuye considerablemente. Además, la evaluación de la eficacia del diccionario fue realizada con entradas aleatorias de otros diccionarios, en lugar de con entradas de usuarios comunes, en lenguaje coloquial, que era su principal objetivo, por lo que no se trata de una evaluación apropiada.

Otro intento de encontrar una forma de obtener un término a partir de su definición es el presentado por Dutoit y Nugues (2002). Este diccionario para el francés toma en cuenta la diferencia entre las palabras de un usuario común para describir un término y aquellas utilizadas por diccionarios formales. Para tratar esta diferencia (buscando la menor diferencia), utiliza una base de datos de palabras organizada en forma jerárquica, de manera que las relaciones de hiponimia e hiperonimia son identificables de forma automática. Este diccionario arroja resultados moderadamente buenos, aunque hay que considerar que su evaluación fue realizada con definiciones que los mismos autores propusieron, por lo que la imparcialidad de las consultas se pone un poco en duda. Además, no se consideran las relaciones sinónimas de las

palabras, por lo que, aunque proporciona al usuario cierta libertad de intercambiar palabras por sus hiperónimos e hipónimos, no considera grupos de sinónimos.

Existe también un diccionario onomasiológico en lenguaje natural desarrollado para el idioma turco por El-Kahlout y Oflazer (2004). En este diccionario se toman en cuenta algunas relaciones sinonímicas entre palabras, así como similitud de definiciones mediante conteo de palabras coincidentes en el mismo orden y en subconjuntos de dichas palabras. Los resultados son moderadamente buenos: 66% de las veces el término buscado por el usuario se encuentra entre los primeros cincuenta términos candidatos, mientras que al utilizar como consulta entradas de otros diccionarios el término buscado se encuentra un 92% de las veces entre los primeros cincuenta términos propuestos. Sin embargo, este diccionario no toma en consideración el uso de palabras coloquiales por parte de los usuarios, además de que el número de términos candidatos considerados en la evaluación es alto (cincuenta); y no se precisa cuál es el posicionamiento promedio del término buscado entre las listas de términos sugeridos.

En inglés, un diccionario onomasiológico en lenguaje natural que arroja resultados aceptables es el OneLook Reverse Dictionary⁷. Este diccionario se encuentra disponible en línea, además de permitir hacer consultas en lenguaje natural, también cuenta con la opción de ingresar expresiones regulares. Desafortunadamente no se especifica cómo funciona internamente el motor de búsqueda del mismo. Únicamente se indica que contiene una base de datos indexada con la información extraída de cientos de diccionarios en línea, enciclopedias y otros sitios de referencia.

Para el español se cuenta con el trabajo realizado por el Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería de la UNAM, el cual trabaja —entre otros proyectos— en la elaboración de diccionarios onomasiológicos electrónicos en línea para diversas áreas temáticas. El motor de búsqueda del Diccionario Electrónico de Búsqueda Onomasiológica (DEBO) del GIL trabaja contemplando relaciones sinonímicas entre palabras, así como un ordenamiento de los términos candidatos basado en el número de coincidencias entre las palabras sinónimas de la consulta y las de las definiciones almacenadas en su base de datos. Buendía *et al.* (2005) reportan a detalle la metodología de búsqueda onomasiológica planteada.

⁷ <http://www.onelook.com>

En esta tesis se propone una mejora para el motor del búsqueda del DEBO del GIL. Con este diccionario se tienen las ventajas de que se cuenta con la descripción completa de su estructura y algoritmo de búsqueda, además de estar enfocado al idioma español.

Capítulo 2. Diccionario Electrónico de Búsqueda

Onomasiológica, DEBO

El DEBO es el resultado de una investigación doctoral de Sierra (1999) realizada en el Instituto de Ciencia y Tecnología de la Universidad de Manchester (UMIST) en el área de ingeniería lingüística. Posteriormente, cuatro estudiantes de licenciatura (Buendía *et al.*, 2005) —dos de Ingeniería en Computación y dos de la Licenciatura en Lengua y Literaturas Hispánicas— de la UNAM realizaron un trabajo de tesis en el que diseñaron un diccionario en línea basándose en la investigación de Sierra (1999). El presente trabajo se basa en el diccionario diseñado por Buendía *et al.* (2005) y su objetivo principal consiste en desarrollar una metodología para semi-automatizar el proceso de creación de diccionarios onomasiológicos, además de mejorar el motor de búsqueda para obtener resultados más acertados.

2.1. Arquitectura general

De acuerdo con Sierra (2007), el DEBO es un sistema integral que agrupa tres módulos principales (ver Figura 5): la *interfaz*, el *sistema de búsqueda onomasiológica* y la *base de conocimientos léxicos*. El primero se trata de un sistema de intercambio de información con el usuario donde éste introduce la descripción del concepto que busca en lenguaje natural. El segundo módulo consiste en el motor de búsqueda que permite asociar términos y definiciones con las palabras clave extraídas de la consulta del usuario. Finalmente, el tercer módulo corresponde a los términos, definiciones, palabras clave y paradigmas semánticos, es decir, agrupa todo el conocimiento en el que el motor de búsqueda se basa para dar una respuesta al usuario. Este trabajo se encuentra enfocado a la semi-automatización de la obtención y adecuación de la información del tercer módulo, así como a la mejora en el desempeño del módulo de búsqueda. Ambos módulos son detallados a continuación.

2.1.1. La base de conocimientos del diccionario

Para que el diccionario onomasiológico pueda dar una respuesta al usuario, es necesario que cuente con una serie de *términos* de entre los cuales pueda elegir como candidatos para

responder a la consulta del usuario, además de las *definiciones* de dichos términos con las cuales el sistema pueda comparar con dicha consulta y así determinar qué términos tienen definiciones más similares a las proporcionadas por el usuario. Asimismo, el diccionario debe contar con una serie de *paradigmas* semánticos que permitan al usuario expresarse utilizando su propio léxico, sin preocuparse de dar exactamente las mismas palabras con las que encontraría la definición en un diccionario. Más adelante se detallarán estos elementos con mayor profundidad. Finalmente, el sistema requiere una *lista de paro* para discernir entre palabras clave y palabras funcionales para lograr extraer la información más importante contenida en la definición del usuario. Estos elementos se detallan a continuación.

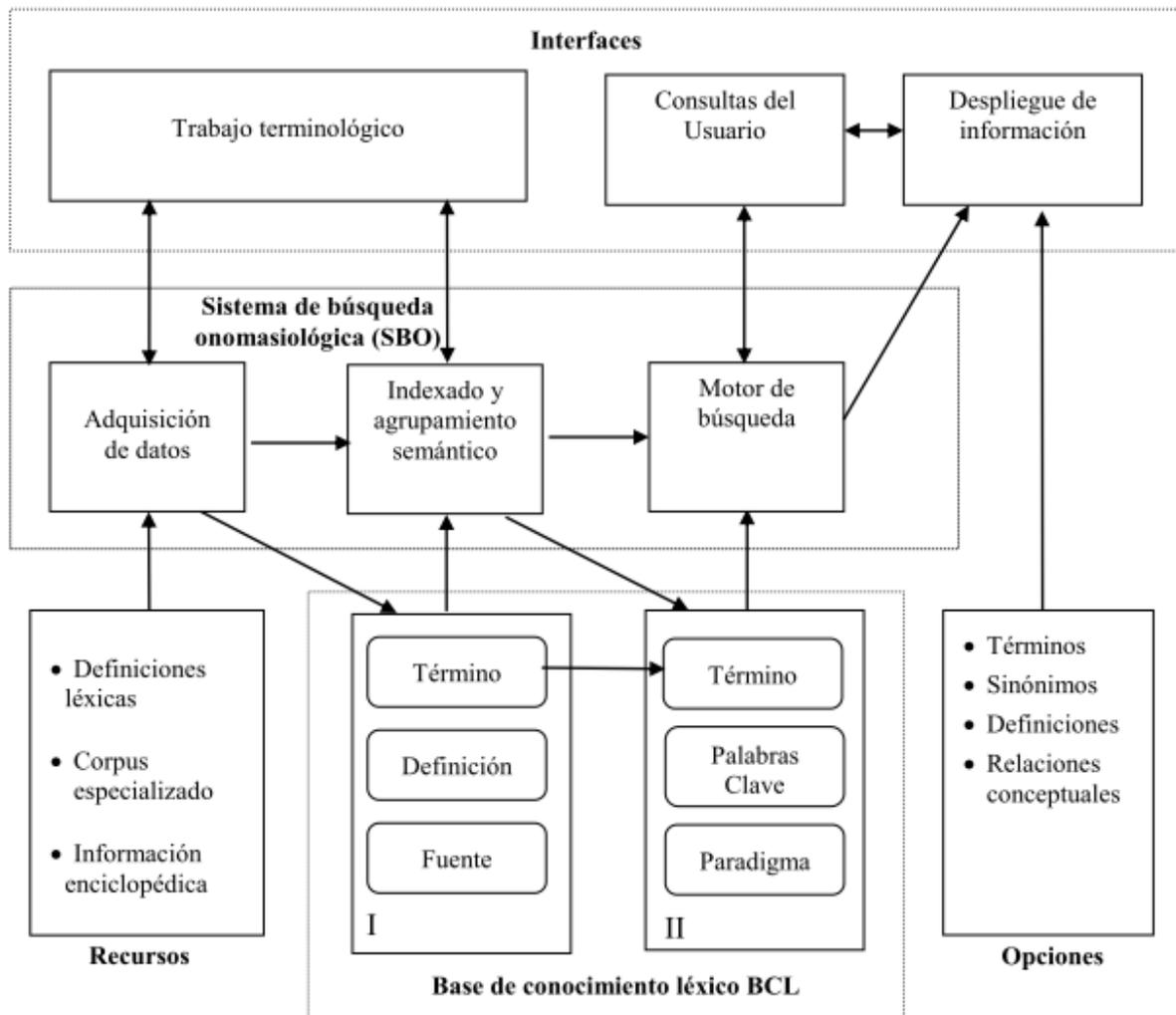


Figura 5: Arquitectura general de DEBO

2.1.1.1. Términos

Se trata de aquellos términos correspondientes al área temática sobre la cual se especializará el diccionario, es decir, los términos de entre los cuales el diccionario buscará para dar respuesta a la consulta del usuario. Estos términos pueden ser seleccionados manualmente por alguna persona experta en terminología o con conocimientos sobre el área de especialidad del diccionario onomasiológico, seleccionados de algún diccionario o texto de especialidad u obtenidos con ayuda de alguna herramienta electrónica que permita su extracción, como la propuesta por Lázaro (2005).

Los términos que se pueden encontrar en el diccionario onomasiológico pueden ser simples (constituidos por una única palabra) o complejos (conformados por dos o más palabras).

2.1.1.2. Definiciones

Las definiciones son parte fundamental de la base de conocimientos del diccionario onomasiológico, pues es el conocimiento a partir del cual el diccionario logra discernir entre los términos y elegir aquél cuya definición se ajuste más a la consulta del usuario.

Las definiciones pueden provenir de diversas fuentes, aunque, como exponen Sierra y Hernández (2012), lo más conveniente es tomar en consideración varias acepciones para cada término, intentado considerar definiciones tanto normativas como no normativas. Las definiciones pueden ser adquiridas de manera manual o mediante el uso de algún sistema automático de extracción de definiciones, como Describe®, del cual se hablará a detalle en el capítulo 4.

2.1.1.3. Paradigmas

El DEBO otorga a sus usuarios la libertad de escribir definiciones utilizando su propio léxico, es decir, no es necesario que los usuarios utilicen las mismas palabras que las usadas en el diccionario para definir un término, basta con que empleen alguna palabra sinónima o con alguna otra fuerte relación semántica en el área de especialización. Para poder tener esta característica en el diccionario onomasiológico, es necesario contar con grupos de palabras con este tipo de relaciones semánticas dentro de la base de conocimientos. A estos grupos de palabras se les conoce como paradigmas.

Un ejemplo de la utilización de los paradigmas es el que muestran Sierra y Hernández (2012) con respecto a dos definiciones de la palabra *barómetro*. La primera definición corresponde a

una posible entrada de un usuario y la segunda corresponde a una definición que podría encontrarse dentro de la base de conocimientos del diccionario onomasiológico:

- *Un objeto que estima cambios en la presión del aire.*
- *Dispositivo utilizado para medir la presión atmosférica.*

Dadas ambas definiciones, un ser humano podría fácilmente hacer la asociación entre ambas definiciones y ver que las dos expresan el mismo sentido aunque con palabras diferentes. La implementación de paradigmas en la base de conocimientos del diccionario podría ayudar a llevar a cabo la misma asociación si contara con los siguientes paradigmas, que corresponden a grupos de sinónimos de WordReference:

Paradigma 1	Paradigma 2	Paradigma 3
objeto dispositivo cosa algo artículo	medir estimar calcular mesurar determinar	aire ambiente atmósfera oxígeno espacio

Tabla 1. Posibles paradigmas obtenidos del diccionario de sinónimos de WordReference

Lo que el proceso de búsqueda del diccionario debe hacer es identificar a qué paradigmas pertenece cada palabra clave y comparar, no palabra a palabra, sino paradigma a paradigma. Con ello, ambas consultas quedarían con una forma similar a la siguiente.

- **Paradigma_1 + Paradigma_2 + paradigma_y + paradigma_z + Paradigma_3.**
- **Paradigma_1 + paradigma_x + Paradigma_2 + paradigma_z + Paradigma_3.**

Donde paradigma_x corresponde al paradigma en el que se encuentra la palabra *utilizado*, paradigma_y es el paradigma de *cambios* y paradigma_z es el grupo donde se localiza *presión*. Con lo anterior, el sistema debe identificar automáticamente la existencia de una alta correspondencia entre ambas definiciones, pues coinciden en cuatro de los cinco paradigmas que las conformaban.

2.1.1.4. Lista de Paro

Para que el diccionario onomasiológico identifique las palabras clave, es decir, las palabras con contenido, es necesario que tenga un listado de aquellas palabras que en Procesamiento de Lenguaje Natural son consideradas como *vacías*, es decir, las palabras funcionales como *el, la, y, con, como, que, por*, etc. A este listado de palabras funcionales se le conoce como lista de paro.

El diccionario busca cada una de las palabras de las definiciones del usuario y de su base de conocimiento en la lista de paro. Si la palabra se encuentra en la lista de paro, ésta se descarta, de lo contrario, se toma en cuenta para realizar la comparación entre la definición del usuario y las definiciones de la base de conocimientos.

La lista de paro puede generarse manual o automáticamente con algún programa computacional, como *Wordsmith* u obtenerse de forma gratuita de algún sitio público que brinde herramientas para el Procesamiento de Lenguaje Natural. También puede generarse manualmente con el conocimiento que el desarrollador cuenta sobre el español.

2.1.2. Descripción de la base de datos

Los cuatro elementos detallados anteriormente —términos, definiciones, paradigmas y lista de paro— consisten en conocimiento escrito en lenguaje natural y deben ser modificados para ser ingresados y consultados en una base de datos computacional, recibiendo, en algunos casos, cierto tratamiento —del cual se hablará más adelante— para mantener sus relaciones durante las diferentes consultas del usuario de forma tal que las consultas realizadas generen respuestas rápidas y congruentes por parte del sistema. A continuación se detallan las tablas que conforman la base de datos del DEBO y se muestra el diagrama general de la base de datos, donde se pueden apreciar las relaciones entre estas tablas.

2.1.2.1. Tablas

La base de datos del diccionario onomasiológico está compuesta por seis tablas: *terminos_diccionario*, *terminos_generales*, *definiciones*, *paradigmas*, *lista_paro* y *definicion_termino*, interrelacionadas de manera tal que sus elementos mantengan las relaciones y el sentido que tienen en el lenguaje natural.

- *terminos_diccionario*

El objetivo de esta tabla consiste en almacenar los términos del área de especialidad del diccionario, es decir, los términos de entre los cuales el usuario deberá obtener alguna respuesta. El primer campo, *id_termino_d*, corresponde al identificador único para cada término. *Nombre_termino* contiene el nombre completo del término, escrito con su correcta ortografía, mientras que *nombre_termino_lematizado* contiene el nombre del término sin acentos ni caracteres especiales y en su forma lematizada, es decir, cada palabra que lo conforma se encuentra en su forma básica o raíz que lo identifica con su familia (por ejemplo, si la palabra es un verbo, éste se encuentra en infinitivo; en caso de tratarse de un sustantivo, se muestra en su forma singular, mientras que un adjetivo se encontraría en su forma masculina y singular). A continuación se muestra la estructura de esta tabla, así como unos ejemplos de posibles registros.

terminos_diccionario		
	Nombre de columna	Tipo de datos
	<i>id_termino_d</i>	ENTERO
	<i>nombre_termino</i>	TEXTO
	<i>nombre_termino_lematizado</i>	TEXTO

Tabla 2. Estructura de la tabla *terminos_diccionario*

id_termino_d	nombre_termino	nombre_termino_lematizado
1	<i>homosexual</i>	<i>homosexual</i>
2	<i>testículos</i>	<i>testículo</i>
3	<i>sífilis</i>	<i>sífilis</i>
4	<i>características sexuales secundarias</i>	<i>caracteristica sexual secundario</i>

Tabla 3. Ejemplos de tuplas que podrían encontrarse en la tabla *terminos_diccionario*

- *terminos_generales*

Esta tabla se compone del vocabulario general que el sistema puede “comprender”, es decir, todas aquellas palabras que conforman el vocabulario del diccionario onomasiológico. Estos vocablos provienen de la base de conocimientos del diccionario y son los que conforman tanto los términos del diccionario, como las definiciones con las que se cuenta, así como los paradigmas. Todos los vocablos que se encuentran almacenados en esta tabla corresponden exclusivamente a cada una de las diferentes palabras de la base de conocimientos, por lo que el sistema no cuenta con un vocabulario más amplio que éste. El campo *id_termino_g* es el identificador único de cada palabra, mientras que *nombre_termino_lematizado* corresponde al nombre de la palabra en forma lematizada. La estructura y algunas tuplas de ejemplo se muestran a continuación.

terminos_generales		
	Nombre de columna	Tipo de datos
	<i>id_termino_g</i>	ENTERO
	<i>nombre_termino_lematizado</i>	TEXTO

Tabla 4. Estructura de la tabla *terminos_generales*

id_termino_g	nombre_termino_lematizado
1	<i>homosexual</i>
2	<i>persona</i>
3	<i>sentir</i>
4	<i>atraccion</i>
5	<i>sexual</i>
6	<i>mismo</i>
7	<i>sexo</i>

Tabla 5. Ejemplos de tuplas que podrían encontrarse en la tabla *terminos_generales*

- *definiciones*

Ésta es la tabla en donde se resguardan las definiciones de la base de conocimientos del diccionario. Estas definiciones son aquellas contra las que el sistema contrastará las consultas del usuario para determinar su grado de similitud. El campo *id_definicion* corresponde al identificador único de cada una de las definiciones; *id_termino_d* es el campo que hace referencia al término del diccionario que se está definiendo (de la tabla *terminos_diccionario*); *definicion_original* corresponde a la definición del término en su grafía original; *definicion_lematizada* corresponde a la definición en su forma lematizada. La estructura de esta tabla, así como algunos ejemplos de sus registros son los siguientes.

definiciones		
	Nombre de columna	Tipo de datos
	<i>id_definicion</i>	ENTERO
	<i>id_termino_d</i>	ENTERO
	<i>definicion_original</i>	TEXTO
	<i>definicion_lematizada</i>	TEXTO

Tabla 6. Estructura de la tabla *definiciones*

id_definicion	id_termino_d	definicion_original	definicion_lematizada
1	1	<i>Persona que siente atracción sexual por personas de el mismo sexo</i>	<i>persona que sentir atraccion sexual por persona de el mismo sexo</i>
2	2	<i>Son las glándulas sexuales masculinas</i>	<i>son el glandula sexual masculino</i>
4	1	<i>Individuo con cualquier forma de actividad sexual con una persona de el mismo sexo</i>	<i>individuo con cualquiera forma de actividad sexual con uno persona de el mismo sexo</i>

Tabla 7. Ejemplos de tuplas que podrían encontrarse en la tabla *definiciones*

- *paradigmas*

Esta tabla se encarga de la agrupación de palabras con significados similares en grupos llamados paradigmas. Para representar estas agrupaciones, cada grupo tiene un identificador único al que pertenecen las palabras correspondientes al mismo. *id_paradigma* es el campo que contiene el identificador único de los grupos de palabras, mientras que *id_termino_g* corresponde al término general que pertenece al grupo con dicho identificador. A continuación se muestra la arquitectura y algunas tuplas ejemplo de esta tabla.

paradigmas		
	Nombre de columna	Tipo de datos
	<i>id_paradigma</i>	ENTERO
	<i>id_termino_g</i>	ENTERO

Tabla 8. Arquitectura de la tabla *paradigmas*

id_paradigma	id_termino_g
1	1
1	25
1	33
1	44
2	92
2	103
2	133

Tabla 9. Ejemplos de tuplas que podrían encontrarse en la tabla *paradigmas*

- *lista_paro*

Esta tabla contiene el listado de aquellas palabras consideradas como funcionales, es decir, sin contenido. Estas palabras son las que el sistema no tomará en cuenta al momento de hacer comparaciones con las consultas de los usuarios. El único campo con el que cuenta es *palabra*, cuyo contenido en cada registro corresponde a una única palabra funcional sin acentos. La arquitectura de esta tabla y algunos ejemplos de registros de muestran a continuación.

lista_paro		
	Nombre de columna	Tipo de datos
	palabra	TEXTO

Tabla 10. Arquitectura de la tabla *lista_paro*

palabra
<i>a</i>
<i>aca</i>
<i>alguno</i>
<i>asi</i>
<i>aunque</i>
<i>como</i>
<i>o</i>

Tabla 11. Ejemplos de tuplas que podrían encontrarse en la tabla *lista_paro*

- *definicion_termino*

En esta tabla se establece una relación entre la definición y los términos generales con los que ésta fue conformada, es decir, las definiciones de la tabla *definiciones* relacionadas con los términos que la conforman, registrados en la tabla *términos_generales*. Con esta relación es posible identificar qué palabras se utilizan al momento de definir cada término del diccionario y, con ello, realizar una acción comparativa rápida con la entrada del usuario.

definicion_termino		
	Nombre de columna	Tipo de datos
	id_definicion	ENTERO
	id_termino_g	ENTERO

Tabla 12. Arquitectura de la tabla *definicion_termino*

id_definicion	nombre_termino_lematizado
1	2
1	3
1	4
1	5
1	6
1	7
2	2

Tabla 13. Ejemplos de tuplas que podrían encontrarse en la tabla *definicion_termino*

2.1.2.2. Diagrama general de la base de datos

A continuación se muestra el diagrama general de la base de datos que engloba todas las tablas mencionadas anteriormente, así como las relaciones existentes entre éstas.

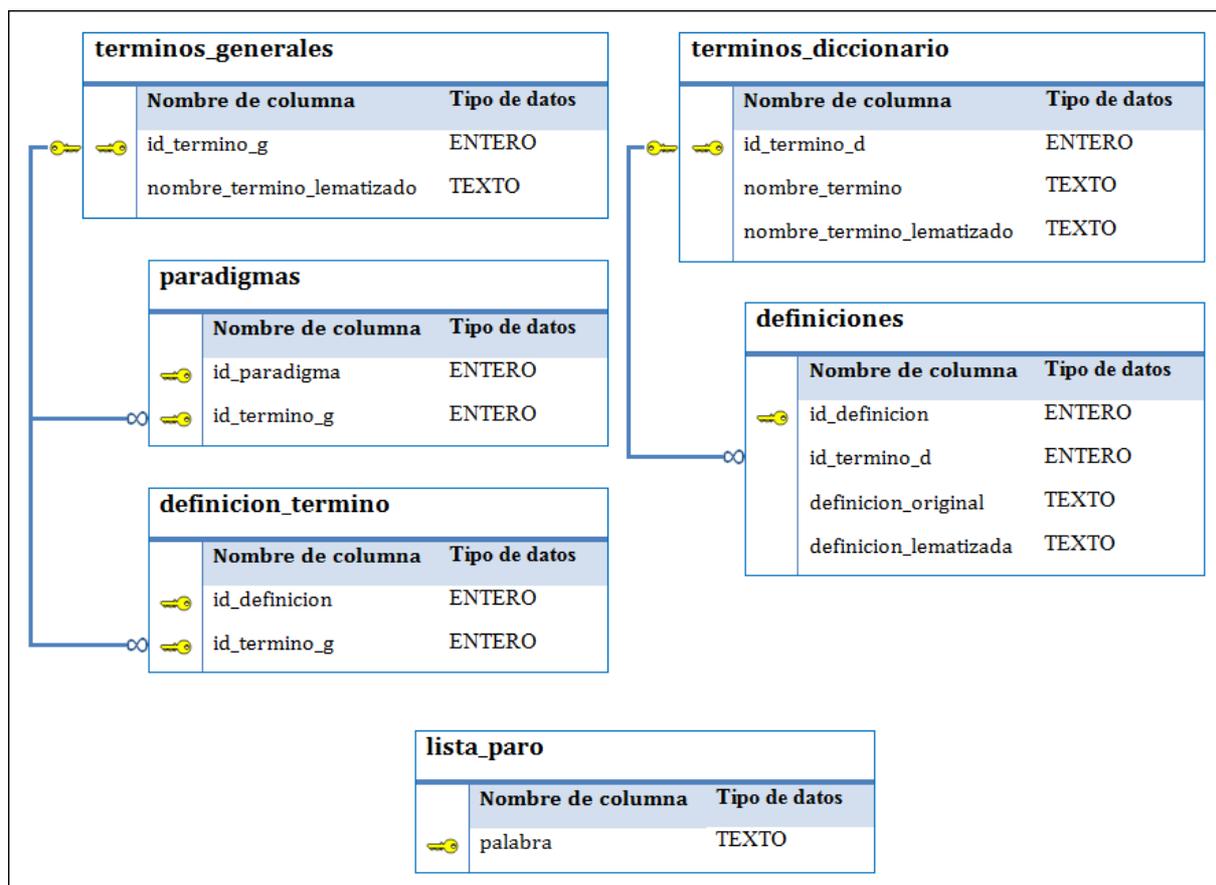


Figura 6. Diagrama general de la base de datos del diccionario onomasiológico

2.2. Algoritmo de búsqueda

El algoritmo de búsqueda del diccionario onomasiológico original contempla hacer asociaciones entre las palabras que escribe el usuario y las que tiene en las definiciones de su base de conocimientos. La forma en la que funciona el motor de búsqueda es relativamente sencilla: el sistema recibe la consulta del usuario y elimina las palabras irrelevantes —utilizando la lista de paro como referencia—, quedándose únicamente con aquellas palabras que son trascendentales en la definición.

A continuación lematiza dichas palabras y detecta los paradigmas de los que estas palabras forman parte. Después busca entre los términos del diccionario aquellos cuyas definiciones tengan asociados los mismos paradigmas encontrados en la consulta del usuario —o la mayor cantidad de estos— y los devuelve como resultado final al usuario, ordenándolos priorizando aquellos que tengan más paradigmas en común, y, en caso de empate, por orden alfabético.

Por ejemplo, suponiendo que se cuenta con los siguientes paradigmas en la base de datos:

rechazo, repudio, odio, aversión, fobia, rabia, rencor, aborrecimiento, desprecio, exclusión	homosexual, lesbiana, gay, maricón, puñal, tortillera, sodomita
Paradigma 1	Paradigma 2

Tabla 14. Ejemplo de paradigmas durante la búsqueda onomasiológica

Y suponiendo que se tienen en la base de datos del diccionario las siguientes definiciones:

Homofobia: *Rechazo a la homosexualidad y a los homosexuales.*

Discriminación: *Cualquier distinción, exclusión o preferencia que tenga por efecto anular o alterar la igualdad de oportunidades o de trato en el empleo.*

Un usuario podría buscar la palabra *homofobia* mediante el siguiente concepto:

Odio a los maricones.

La búsqueda onomasiológica que correspondería a la entrada del usuario se vería como en el siguiente diagrama.

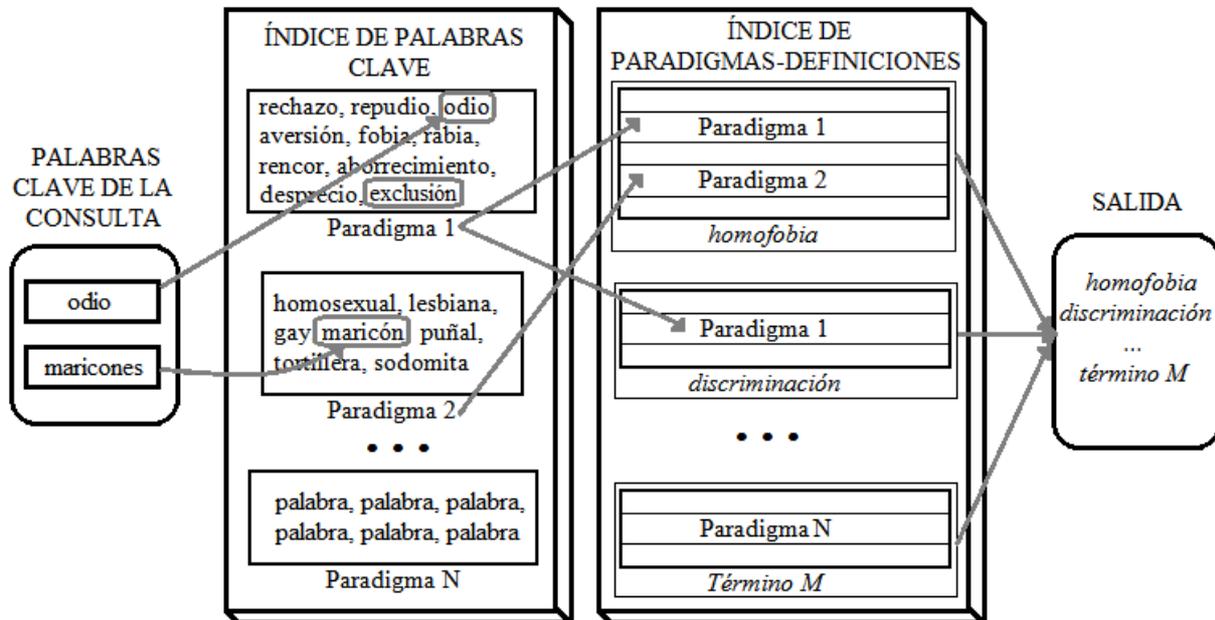


Figura 7. Diagrama del proceso de búsqueda del diccionario onomasiológico

Las únicas palabras clave de la consulta, tras pasar por el filtro de la lista de paro son *odio* y *maricones*, por lo que la consulta del usuario se asocia con dos paradigmas: el **Paradigma 1** y el **Paradigma 2**. Por otro lado, en la base de datos el término *discriminación* se encuentra asociado con el **Paradigma 1** —por la palabra *exclusión*—, mientras que el término *homofobia* se encuentra asociado tanto con el **Paradigma 1** —por la palabra *odio*— como con el **Paradigma 2** —por la palabra *maricón*—. Entonces, al momento de dar una respuesta, el diccionario coloca a la palabra *homofobia* por encima de *discriminación*, pues el número de paradigmas con los que coincide con la entrada del usuario es mayor.

Con este método se obtienen resultados variables y, aunque se suele identificar el término correcto, éste no siempre se encuentra entre los primeros resultados de la lista de salida —que puede llegar a ser larga— y, por ello, es deseable mejorar la efectividad de la búsqueda onomasiológica.

Capítulo 3. Propuesta del motor de búsqueda de diccionarios onomasiológicos

En este capítulo se hablará de una propuesta para mejorar los resultados en la búsqueda onomasiológica del diccionario propuesto originalmente por Sierra (1999) y adaptado por (Buendía *et al.*, 2005). La propuesta debe tomar en cuenta que la velocidad de búsqueda no debe verse afectada negativamente en forma significativa, pues se trata de un sistema que debe dar respuestas en tiempo real a sus usuarios.

La propuesta de mejora contempla dos tipos de cambios: los realizados al algoritmo de búsqueda y aquellos realizados directamente en la estructura de la base de datos.

3.1. Cambios propuestos al algoritmo de búsqueda

El tipo de cambios en el algoritmo de búsqueda contempla modificaciones al método planteado por Sierra (1999), pero sin alterar la idea esencial del diccionario que le permite al usuario expresarse en sus propias palabras de forma libre. Estos cambios se enfocan principalmente en mejorar los resultados de la búsqueda —siendo ésta su principal contribución—, 1) eliminando de la salida términos candidatos irrelevantes y 2) posicionando mejor a los términos meta u objetivo.

3.1.1. Búsqueda en los paradigmas

Al momento de realizar una búsqueda en el diccionario onomasiológico, la consulta del usuario es filtrada y de ella se extraen sus palabras clave. Estas palabras después son buscadas dentro de la tabla de paradigmas para ubicar aquellos paradigmas en los que se encuentran y, de esta manera, buscar de entre los términos aquellos que contengan también dichos paradigmas.

Por ejemplo, si el usuario introdujo la palabra *embarazo*, ésta se podrá encontrar en los siguientes paradigmas obtenidos del diccionario de sinónimos de WordReference.

1. **embarazo**, gestación, preñez, gravidez, preñar, empreñar, fecundar, fertilizar, concebir, engendrar.
2. **embarazo**, impedimento, obstáculo, inconveniente, compromiso, estorbo, molestia, corte.
3. **embarazo**, turbación, apocamiento, cortedad, timidez.
4. preñez, **embarazo**, concepción, gravidez, gestación, vientre, tripa.
5. obstáculo, estorbo, impedimento, inconveniente, óbice, traba, **embarazo**, freno, dificultad, obstrucción.
6. apocamiento, cortedad, timidez, pusilanimidad, ñoñería, flojedad, cobardía, cortedad, vacilación, vergüenza, abatimiento, acoquinamiento, achicamiento, flaqueza, aturdimiento, depresión, desaliento, empacho, encogimiento, humillación, irresolución, languidez, poquedad, retraimiento, turbación, **embarazo**, susto, temor.
7. cortedad, timidez, azoramiento, encogimiento, apocamiento, **embarazo**, vergüenza.
8. gravidez, preñez, gestación, **embarazo**, maternidad, concepción.
9. concebir, imaginar, pensar, entender, percibir, comprender, intuir, inferir, engendrar, fecundar, **embarazar**, procrear.
10. cohibir, refrenar, reprimir, contener, coartar, coercer, limitar, sujetar, cortar, **embarazar**.

Considerando todos los paradigmas en los que la palabra *embarazo* se encuentra —sin importar su posición—, se puede observar que la palabra *embarazo* se relacionará con cualquiera de las siguientes setenta palabras diferentes:

abatimiento, achicamiento, acoquinamiento, apocamiento, aturdimiento, azoramiento, coartar, cobardía, coercer, cohibir, comprender, compromiso, concebir, concepción, contener, cortar, corte, cortedad, depresión, desaliento, dificultad, embarazar, embarazo, empacho, empreñar, encogimiento, engendrar, entender, estorbo, fecundar, fertilizar, flaqueza, flojedad, freno, gestación, gravidez, humillación, imaginar, impedimento, inconveniente, inferir, intuir, irresolución, languidez, limitar, maternidad, molestia, ñoñería, óbice, obstáculo, obstrucción, pensar, percibir, poquedad, preñar, preñez, procrear, pusilanimidad, refrenar, reprimir, retraimiento, sujetar, susto, temor, timidez, traba, tripa, turbación, vacilación, vergüenza y vientre.

Lo que se propone ahora es realizar la búsqueda en la misma forma en la que se suele buscar en un diccionario de sinónimos, en otras palabras, sólo considerar los paradigmas en los que la palabra *embarazo* se encuentre como índice, es decir, como primer elemento, y no considerar el resto de los paradigmas en los que *embarazo* se encuentra listado entre el resto de los términos. Con ello, se podrían tener las palabras que se relacionan por una **correspondencia sinonímica directa**, es decir, aquellas que se encuentran asociadas en un mismo listado en el diccionario de sinónimos. De esta manera, la lista de palabras relacionadas con *embarazo* se reduce a una veintena:

apocamiento, compromiso, concebir, corte, cortedad, embarazo, empreñar, engendrar, estorbo, fecundar, fertilizar, gestación, gravidez, impedimento, inconveniente, molestia, obstáculo, preñar, preñez, timidez y turbación.

Con este cambio en la forma de búsqueda en los paradigmas, las palabras relacionadas disminuyeron para este ejemplo en más de la mitad y, además, se eliminaron palabras como *ñoñería, susto, pensar, imaginar o cortar*, que tienen una relación semántica menor con la palabra *embarazo* y que, por ende, generarían concordancias con términos que no tienen una fuerte relación con el término buscado. Únicamente se conservaron las palabras con asociaciones semánticas en los tres sentidos de la palabra *embarazo*: embarazo visto como impedimento o dificultad, como preñez de la mujer o como falta de soltura o vergüenza.

De aquí se propone entonces determinar en cada grupo de paradigmas el índice del mismo, es decir, la palabra inicial del listado y con la cual son asociadas directamente las palabras de cada grupo. Si los paradigmas fueron extraídos de un diccionario de sinónimos, entonces los índices se encuentran claramente identificados.

3.1.2. Búsqueda con n-paradigmas

Con la metodología de búsqueda en el diccionario onomasiológico planteada originalmente por Sierra (1999) el único criterio de posicionamiento de un término consiste en contar el número de

correspondencias entre los paradigmas asociados a la consulta del usuario y los paradigmas asociados a las definiciones en la base de conocimientos.

Supóngase que se cuenta con las siguientes definiciones en la base de conocimientos:

Pedofilia: atracción sexual hacia niños o adolescentes.

Desarrollo sexual: cambios físicos y psicológicos que suceden a la niñez del ser humano y que transcurre desde la pubertad, con el despertar del deseo sexual, hasta el pleno desarrollo físico.

Además, se cuenta con los siguientes paradigmas:

atracción, seducción, captación, deseo, apetito, antojo, anhelo	sexual, genital, sensual, carnal, erótico	niño, chamaco, criatura, peque, chico, mocososo, chiquillo
Paradigma 1	Paradigma 2	Paradigma 3

Tabla 15. Ejemplo de paradigmas durante la búsqueda onomasiológica

Suponiendo que se tiene la siguiente entrada del usuario:

Pedofilia: deseo sexual por los niños.

La consulta del usuario quedaría asociada a los paradigmas 1 (*deseo*), 2 (*sexual*) y 3 (*niño*). Al revisar la definición en la base de conocimientos de *pedofilia*, ésta se encontrará relacionada —entre otras— con los paradigmas 1 (*atracción*), 2 (*sexual*) y 3 (*niño*), mientras que la definición de *desarrollo sexual* estará asociada con 3 (*niño*), 1 (*deseo*) y 2 (*sexual*).

Con la metodología de búsqueda actual, el sistema ubicaría a *pedofilia* y *desarrollo sexual* en con igual posicionamiento para ser la solución a la consulta del usuario. Más aún, al estar *desarrollo sexual* alfabéticamente antes de *pedofilia*, la primera iría por encima de ésta en la salida del diccionario.

El problema principal que explica este resultado consiste en que la metodología propuesta originalmente no contempla considerar el orden en el que aparecen dichos paradigmas, sino que únicamente toma en consideración el número de coincidencias, por lo que no hay más criterios

para discernir entre los términos propuestos y subir en grado de importancia al término con mayor similitud en la forma de organización de las ideas con las del usuario.

En Procesamiento de Lenguaje Natural existen los *n*-gramas, que consisten en subsecuencias de *n* elementos de una secuencia de gramas o palabras. Por ejemplo, en la oración *el niño tiene hambre*, los posibles 2-gramas son: *el niño*, *niño tiene* y *tiene hambre*, mientras que los posibles 3-gramas son: *el niño tiene* y *niño tiene hambre*. Para esta tesis se propone hacer una analogía con los *n*-gramas, llamada *n*-paradigmas que, como se puede deducir fácilmente, consiste en considerar no sólo coincidencias en la cantidad de paradigmas sino en su orden, priorizando aquellas con el mayor número de subsecuencias posibles o las subsecuencias mayores en las secuencias de paradigmas.

Entonces, retomando el ejemplo anterior, la definición del usuario estaría conformada de la siguiente manera:

Paradigma 1 + Paradigma 2 + Paradigma 3

Los 2-paradigmas posibles en esta consulta son:

- **Paradigma 1 + Paradigma 2**
- **Paradigma 2 + Paradigma 3**

El único 3-paradigma de la consulta es:

- **Paradigma 1 + Paradigma 2 + Paradigma 3**

Ahora, los 2-paradigmas para *pedofilia* son:

- **Paradigma 1 + Paradigma 2**
- **Paradigma 2 + Paradigma 3**
- **Paradigma 3 + Paradigma *n***

Mientras que sus 3-paradigmas son:

- **Paradigma 1 + Paradigma 2 + Paradigma 3**
- **Paradigma 2 + Paradigma 3 + Paradigma *n***

En el caso de *desarrollo sexual*, algunos de sus 2-paradigmas más representativos para este caso son:

- ...
- **Paradigma w + Paradigma 3**
- **Paradigma 3 + Paradigma x**
- ...
- **Paradigma y + Paradigma 1**
- **Paradigma 1 + Paradigma 2**
- **Paradigma 2 + Paradigma z**
- ...

Mientras que sus 3-paradigmas más representativos son:

- ...
- **Paradigma v + Paradigma w + Paradigma 3**
- **Paradigma w + Paradigma 3 + Paradigma x**
- **Paradigma 3 + Paradigma x + Paradigma m**
- ...
- **Paradigma n + Paradigma y + Paradigma 1**
- **Paradigma y + Paradigma 1 + Paradigma 2**
- **Paradigma 1 + Paradigma 2 + Paradigma z**
- **Paradigma 2 + Paradigma z + + Paradigma p**
- ...

De lo anterior se puede observar que, aunque *deseo sexual* y *pedofilia* coinciden en el número de paradigmas asociados a la consulta del usuario, existe una correspondencia mayor entre la consulta del usuario y la definición de *pedofilia*, pues no correspondieron únicamente en paradigmas individuales o 2-paradigmas, sino inclusive en grupos de hasta 3-paradigmas, por lo que, siguiendo esta pauta, *pedofilia* quedaría posicionado mejor en el ranking de salida del diccionario.

Como se puede ver, para extraer los n-paradigmas y hacer comparaciones se deben realizar muchas operaciones y el número de combinaciones de palabras que se generan es alto. Por ello, decidí únicamente considerar 2-paradigmas y 3-paradigmas, con la finalidad de no incrementar el tiempo de cómputo requerido para la búsqueda, además de que, al tratarse de definiciones, las consultas de los usuarios no suelen contener demasiadas palabras.

La forma en la que se ponderan las definiciones y se sube de categoría a los términos mostrados al usuario es mediante puntuaciones que permiten destacar visiblemente a aquellos términos que coincidieron en 3-paradigmas sobre 2-paradigmas y estos últimos sobre paradigmas simples. Para ello, por cada 3-paradigma con el que coincida un término, se suman 100 puntos; un 2-paradigma suma 10 y un paradigma sencillo suma únicamente 1. De esta manera, aquellos términos resultantes que no sólo hayan coincidido en paradigmas, sino que además de hayan ingresado en un orden similar al de la consulta del usuario tienen clara ventaja sobre el resto.

3.2. Cambios propuestos a la base de datos

Los cambios que se proponen para la base de datos están orientados a disminuir el tiempo de procesamiento durante el algoritmo de búsqueda del diccionario, así como a adaptarla a los cambios propuestos para el algoritmo mencionado en 3.1.

3.2.1. Indización

Para agilizar la búsqueda en el diccionario cuando se cuenta con un gran número de definiciones, paradigmas o cualesquiera elementos en la base de conocimientos, es conveniente realizar una tabla índice que asocie de forma automática a los términos con los paradigmas, pues actualmente para llevar a cabo dicha asociación es necesario dar un tratamiento particular a las definiciones (eliminar las palabras funcionales con ayuda de la lista de paro), para posteriormente asociarlas con los paradigmas (realizar consultas palabra a palabra) y, finalmente, reasignarlas al término de la definición.

Esta estrategia de búsqueda resulta práctica para diccionarios con pocos términos o definiciones, así como para cuando se tienen pocos paradigmas. Sin embargo, lo más común es que los diccionarios contengan una gran cantidad de términos especializados y, si los paradigmas se extraen de un diccionario de sinónimos, es bastante probable que el tamaño de esta tabla sea grande y las comparaciones palabra a palabra en cada consulta sean costosas. Por lo mismo, conviene tener una tabla que relacione directamente estos elementos. La tabla propuesta es: *terminos_diccionario_paradigma*, y su estructura se muestra a continuación.

terminos_diccionario_paradigma		
	Nombre de columna	Tipo de datos
	id_termino_d	ENTERO
	id_paradigma	ENTERO

Tabla 16. Arquitectura de la tabla *terminos_diccionario_paradigma*

Para ejemplificar los datos que podría contener esta tabla, considere que se cuenta con las siguientes definiciones en la base de conocimientos para el término *bisexual* con identificador 7.

- ***bisexual***: Persona que tiene relaciones sexuales tanto con hombres como con mujeres.
- ***bisexual***: Persona que siente atracción sexual hacia individuos de ambos sexos

Y que se cuenta también con los siguientes paradigmas:

persona, ser, individuo, sujeto	tener, poseer, conservar, guardar	nexo, ligazón, relación, vínculo	sexual, genital, sensual, carnal, sexo, erótico	tanto
Paradigma 1	Paradigma 2	Paradigma 3	Paradigma 4	Paradigma 5
hombre, sujeto, varón, macho, masculino	mujer, hembra, fémica, dama, señora, señorita	sentir, palpar, experimentar, percibir	atracción, deseo, seducción	ambos, dos
Paradigma 6	Paradigma 7	Paradigma 8	Paradigma 9	Paradigma 10

Tabla 17. Ejemplo de paradigmas durante la búsqueda onomasiológica

Los valores en la tabla *terminos_diccionario_paradigma* que representarían la asociación entre el término *bisexual* y los paradigmas anteriores se muestran en la Tabla 18.

id_termino_d	id_paradigma
7	1
7	2
7	3
7	4
7	5
7	6
7	7
7	8
7	9
7	10

Tabla 18. Ejemplos de tuplas que podrían encontrarse en la tabla *terminos_diccionario_paradigma*

Con esta tabla adicional, la búsqueda de paradigmas coincidentes sería efectuada de manera mucho más directa y rápida.

3.2.2. Consideración de n-paradigmas

Para tomar en cuenta a los n-paradigmas es necesario considerar no solamente a los paradigmas con los que se asocia un término sino, además, el orden en el que éstos aparecen. De ahí que se decidió agregar el campo *orden_en_definicion*, para registrar la posición en la que aparece el paradigma y eliminar de la tabla *terminos_diccionario_paradigma* las claves primarias, pues un mismo término puede tener asociados dos paradigmas idénticos que aparecen en diferentes definiciones pero en la misma posición. La estructura de la tabla *terminos_diccionario_paradigma* resultante se muestra a continuación.

terminos_diccionario_paradigma	
Nombre de columna	Tipo de datos
id_termino_d	ENTERO
id_paradigma	ENTERO
orden_en_definicion	ENTERO

Tabla 19. Estructura de la tabla *terminos_diccionario_paradigma*

Para clarificar su funcionamiento, supóngase que se tienen las mismas definiciones y ejemplos que los mencionados en 3.2.1. Los valores en la tabla que representarían la asociación entre el término *bisexual* y los paradigmas, en su respectivo orden son los siguientes.

id_termino_d	id_paradigma	orden_en_definicion
7	1	1
7	2	2
7	3	3
7	4	4
7	5	5
7	6	6
7	7	7
7	1	1
7	8	2
7	9	3
7	4	4
7	1	5
7	10	6
7	4	7

Tabla 20. Ejemplos de tuplas que podrían encontrarse en la tabla *terminos_diccionario_paradigma* con la consideración del orden de los paradigmas

El diagrama general de la base de datos resultante tras estas modificaciones es el mostrado en la Figura 8.

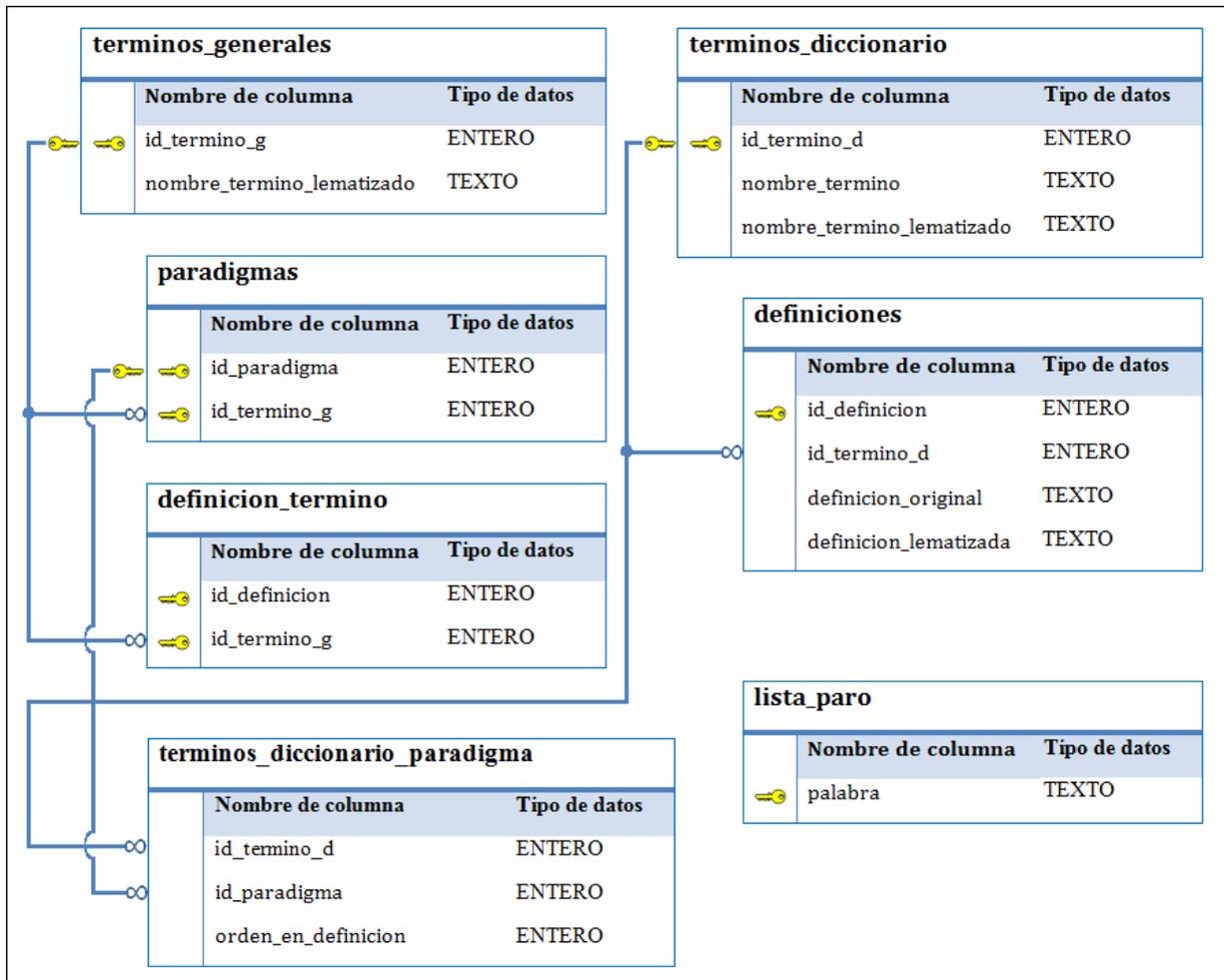


Figura 8. Diagrama general de la base de datos final

Capítulo 4. Generación semi-automática de la base de datos de diccionarios onomasiológicos

En su libro, Gelbukh y Sidorov (2006) exponen la complejidad que representa el trabajo en procesamiento de lenguaje natural (PLN) particularmente para el idioma español. Se hace referencia a las diferentes técnicas de PLN para generar recursos léxicos para el área, así como la aplicación de recursos léxicos —particularmente diccionarios— en otras técnicas de PLN y el manejo de grandes volúmenes de información. Esta tesis propone una metodología para crear desde cero y de forma semi-automática diccionarios onomasiológicos especializados que funcionen con la estrategia de búsqueda y estructura planteadas por Sierra (1999). Esta metodología abarca desde la obtención de los términos y definiciones de la especialidad del diccionario, hasta el tratamiento automatizado de éstos, con el objetivo de dejarlos en un formato adecuado para ser ingresados a una base de datos generada automáticamente con este propósito específico.

Para la realización de esta tesis se desarrolló el Sistema para la Generación Automática de Diccionarios Onomasiológicos (SGADO), cuya función consiste en crear de forma automática una base de datos con la configuración propuesta en el capítulo 3, así como el poblado de la base de conocimiento con la que se deberá alimentar el diccionario. A continuación se habla a detalle de SGADO.

4.1. Entradas

En 2.1.1 se habló de los elementos fundamentales que contiene la base de conocimientos del diccionario, los cuales son: términos, definiciones, paradigmas y lista de paro. Todos estos elementos constituyen las entradas que SGADO recibirá e ingresará en la base de datos especializada.

4.1.1. Términos

Existen diferentes métodos para la extracción automática de términos. Cabrera (2011) hace una investigación y detalla los esfuerzos que actualmente se realizan en el área de Procesamiento de Lenguaje Natural para llevar a cabo en forma automática la extracción terminológica. Como él expone, existen y se siguen generando diversas metodologías y tipos de sistemas de extracción terminológica, que se dividen en los sistemas basados en conocimiento lingüístico, los basados en conocimiento estadístico y los híbridos, de se basan en ambos conocimientos.

Lázaro (2005) propone una metodología basada en conocimiento estadístico para extraer términos de un área especializada mediante el uso de *Wordsmith*⁸, una herramienta de software para análisis léxico que cuenta, entre otras funcionalidades, con la opción de presentar una lista de palabras generales de textos, además de tener la herramienta *keywords*, que arroja, a partir de la lista de palabras generales, una lista de palabras clave que son muy buenos candidatos a términos. Esta metodología se eligió debido a que es que fue la utilizada para extraer los términos del diccionario de sexualidad con el que se evaluó esta tesis.

La forma en la que se utiliza *Wordsmith* para obtener términos de un área específica consiste en proporcionar una serie de textos que hablan sobre temas generales, así como textos sobre el área particular de interés. Una vez que el sistema cuenta con ambos tipos de textos, la herramienta *keywords* determina cuáles son aquellos vocablos que aparecen con mayor frecuencia en los textos de especialidad pero que no aparecen con tanta frecuencia en los textos genéricos. Estos vocablos son los que con mayor seguridad corresponden a palabras clave y a posibles términos del área de especialidad, pues se trata de vocablos que son muy utilizados en el área pero no en el léxico común, lo que sugiere cierto nivel de especialización del usuario.

La idea para la extracción de los términos especializados consiste en proporcionarle a *Wordsmith* una serie de textos de la especialidad del diccionario onomasiológico, así como textos generales y hacer que éste, mediante su herramienta *keywords*, entregue un listado de palabras clave que

⁸ <http://www.lexically.net/wordsmith/>

son candidatos a término. De esta lista se pueden seleccionar manualmente los términos que se definirán en el diccionario.

4.1.2. Definiciones

Como se había mencionado previamente, las definiciones constituyen la parte esencial de la base de conocimientos del diccionario, pues en ellas se basará el sistema para realizar las comparaciones de las entradas de los usuarios. Es importante que las definiciones del diccionario sean ricas y variadas, pues esto permitirá que el diccionario tenga más información sobre cada término y sus diversas acepciones y formas de definirlo. En el Grupo de Ingeniería Lingüística existe una herramienta llamada Describe®, la cual fue desarrollada dentro del grupo y consiste en un sistema que busca automáticamente candidatos a definiciones de términos particulares en Internet, lo que permite obtener definiciones que provienen no sólo de hablantes expertos en el área, sino de la sociedad en general.

4.1.2.1. Describe®

Describe® es la herramienta ideal para la extracción de definiciones para el diccionario onomasiológico (Sierra et al., 2009), pues el tener definiciones tanto normativas como no normativas permite que la base de conocimientos del diccionario sea rica en construcciones y expresiones definatorias, así como en ideas y modos formales e informales de expresión. Esta riqueza es importante, pues los usuarios para los que está contemplado el diccionario no son necesariamente expertos en el área de especialidad del diccionario.

Como exponen Alarcón, Bach y Sierra (2008), la manera en la que Describe® extrae los candidatos a definiciones es mediante la utilización de ECODE, otra herramienta desarrollada en el GIL y que mediante patrones sintácticos busca contextos definatorios, es decir, los contextos discursivos donde la información relevante utilizada para definir un término puede ser encontrada (Alarcón y Sierra, 2003).

Los patrones definatorios que contempla ECODE son verbos definatorios, como *definir* y *significar*, marcadores discursivos como *es decir* y *o sea*, además de patrones tipográficos, como comas y paréntesis y patrones pragmáticos como *en términos generales* o *en este sentido*.

Existen quince patrones definatorios verbales que ECODE toma en consideración al momento de buscar definiciones. Estos patrones definatorios fueron extraídos por Alarcón *et al.* de un grupo de textos (corpus) del Instituto Universitario de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra. Este corpus está compuesto por textos especializados en las áreas de economía, medio ambiente, leyes, genética, medicina, informática y lenguaje general.

Patrones verbales definatorios			
• <i>concebir</i>	• <i>significar</i>	• <i>denominar también</i>	• <i>usar para</i>
• <i>definir</i>	• <i>consistir de</i>	• <i>llamar también</i>	• <i>utilizar como</i>
• <i>entender</i>	• <i>consistir en</i>	• <i>servir para</i>	• <i>utilizar para</i>
• <i>identificar</i>	• <i>constar de</i>	• <i>usar como</i>	

Tabla 21. Patrones definatorios verbales utilizados por ECODE

Tras la identificación de los patrones verbales, ECODE realiza un proceso automático de etiquetado que le permite identificar las partes del contexto definatorio, tales como el término definido, la definición, el patrón verbal utilizado y aquellos patrones pragmáticos que pudieran acompañar a la definición. Para ejemplificar el funcionamiento de ECODE, se presenta un ejemplo de contexto definatorio, así como de los elementos que ECODE puede diferenciar y separar.

En términos médicos el aborto se define como la interrupción del embarazo antes de la viabilidad del feto.

Término: *aborto*

Definición: *interrupción del embarazo antes de la viabilidad del feto.*

Patrón verbal conector: *definir (como)*

Patrones pragmáticos como modificadores del contexto: *En términos médicos*

Sierra y Hernández (2012) explican con mayor detalle el funcionamiento de esta herramienta, así como la metodología que utilizan para la extracción de candidatos a definiciones.

4.1.2.2. Sistema para la selección y depuración de los resultados de Describe®

La salida de Describe® consiste en una serie de candidatos a definición del término que se le solicite, por lo que es necesario depurar esta lista de tal manera que únicamente se conserven aquellos candidatos que sean definiciones. En esta tesis se desarrolló —en PHP con PostgreSQL— un sistema para depurar los resultados de Describe® de una manera cómoda y sencilla.



Figura 9. Pantalla principal del sistema para la selección y depuración de los resultados de Describe®

Este sistema permite ver un listado con los candidatos a definición proporcionados por Describe® y seleccionar manualmente de aquí aquellos que se considere que realmente son definiciones. Además, el sistema permite determinar —con un propósito de investigación lingüístico— el tipo de definición del cual se trata. La Figura 10 muestra la información desplegada para los candidatos a definición obtenidos para el término *travestismo*.

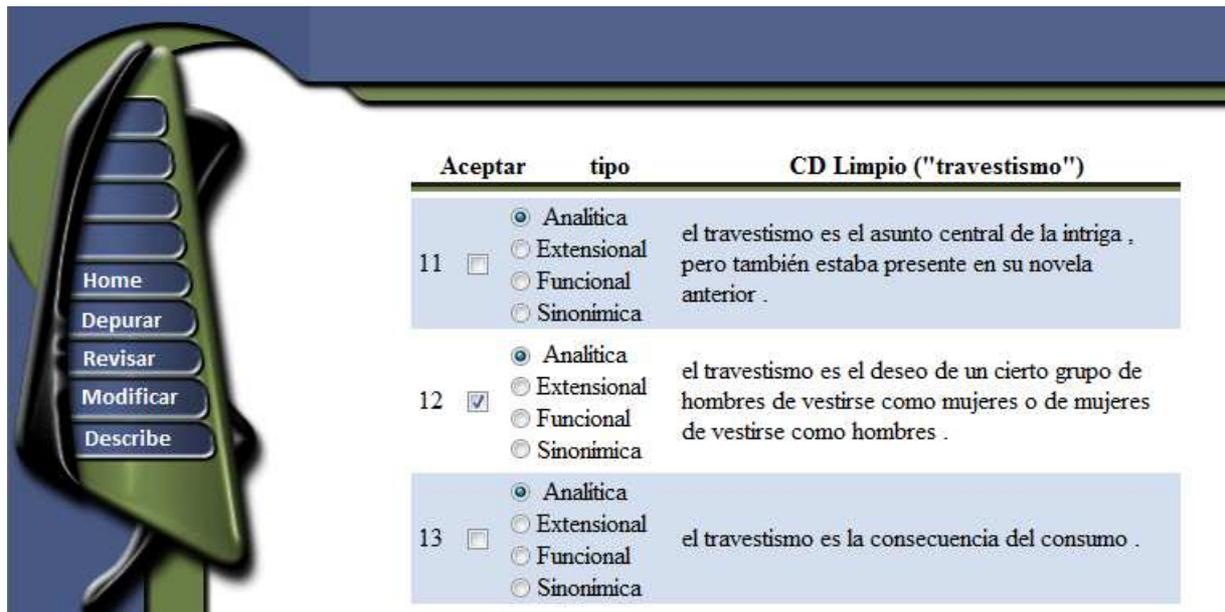


Figura 10. Pantalla de selección de candidatos a definición para el término *travestismo*

4.1.3. Paradigmas

Un método fácil, rápido y automático para obtener los paradigmas consiste en utilizar los datos contenidos en algún diccionario de sinónimos electrónico. Estos diccionarios ya contienen las palabras agrupadas por su alta similitud semántica en sus varias acepciones, además de que la estructura en la que presentan sus datos se adapta perfectamente a la metodología de búsqueda en paradigmas propuesta en el capítulo 3.1.1.

Es importante considerar que en ciertas áreas, como sexualidad, conviene que el diccionario sea capaz de asociar palabras que en ciertos contextos no normativos son similares semánticamente, es decir, palabras coloquiales que las personas utilizan en el contexto en particular para referirse a términos sinónimos. Por ejemplo, la palabra *pilín* es una palabra que un niño mexicano puede utilizar para referirse al pene. Ante estos casos tal vez sea conveniente agregar manualmente algunas palabras dentro de los grupos de sinónimos.

4.1.4. Lista de paro

Existen disponibles en Internet recursos para el área de Procesamiento de Lenguaje Natural que incluyen listas de paro en diversos idiomas. Estas listas contienen aquellas palabras que en un

lenguaje general se pueden considerar como vacías. Se puede tomar una de estas listas que no suelen ser demasiado grandes y dejarla tal cual o depurarla manualmente dependiendo del tema, pues algunas palabras, aunque en lenguaje general no tienen mucha relevancia, para ciertas áreas o en ciertos contextos resultan ser palabras características.

También se pueden utilizar herramientas electrónicas como *Wordsmith* de la siguiente forma: se ingresa una serie de textos de lenguaje general y se extrae el listado de palabras generales que va ordenado por el número de apariciones de dicha palabra. Las palabras que se repiten más suelen ser las palabras funcionales, por lo que esta lista es una buena guía para obtener las palabras de la lista de paro.

4.2. Tratamiento automático de las entradas

Como se mencionó en el capítulo 2.1.2.1, es necesario que las palabras se encuentren lematizadas, es decir, que todos los elementos de una misma familia se pasen a una forma única para hacerles referencia. Por ejemplo, al lematizar las palabras *come*, *comió*, *comiste*, *comerá*, *comeremos*, *comería* y *comérselo*, todas remiten en su forma lematizada a la palabra *comer*. Existen diversos programas que se pueden utilizar para realizar esta tarea de forma automática, sin embargo, para el desarrollo de esta metodología se decidió utilizar FreeLing⁹ (Carreras *et al.* (2004), Atserias *et al.* (2006) y Padró *et al.* (2010)), una suite de analizadores de lenguajes en software libre.

FreeLing fue elegido de entre otras herramientas porque constituye una herramienta de Procesamientos de Lenguaje Natural con un gran contenido para el español, además de que actualmente se encuentra bajo desarrollo un módulo de corrección ortográfica que podría ser utilizado más adelante en el diccionario onomasiológico con altos beneficios, pues los usuarios suelen cometer errores ortográficos que impiden al sistema detectar correctamente la palabra que pretendían ingresar.

⁹ <http://nlp.lsi.upc.edu/freeling/>

Para realizar el tratamiento automático de las entradas, se desarrollaron en SGADO algunos shell scripts que reciben las entradas (términos, definiciones, paradigmas y lista de paro) en un formato particular y SGADO ejecuta FreeLing con las opciones requeridas para que su salida sea la óptima al momento de realizar el ingreso de los datos a la base de datos. A continuación se describe la manera en la que los shell scripts de SGADO reciben las entradas.

4.2.1. Términos y definiciones

Las definiciones deberán encontrarse en un formato de texto plano con una codificación de caracteres tipo ISO 8859-1. Cada término irá definido en un archivo individual, y en este archivo la primera línea deberá corresponder al nombre del término (éste puede ser multipalabra). Cada definición deberá ir en una línea distinta, y al final de cada línea se deberá colocar el marcador con un tabulador previo “_____”. Este marcador permite identificar el final de una definición, pues algunas pueden abarcar varias oraciones. La generación de estos archivos puede ser automatizada fácilmente mediante una extracción automática de los elementos seleccionados mediante el Sistema para la selección y depuración de los resultados de Describe®.

Por ejemplo, un archivo para el término *violación sexual* con 3 definiciones contendría la información con el siguiente formato:

```
1 violación sexual _____
2 Acto sexual forzado ya sea físicamente, emocionalmente o por coacción verbal. _____
3 es el acto sexual perpetrado con una persona tomandola por la fuerza. _____
4 La violación se define como la penetración sexual por la fuerza. _____
```

Figura 11. Ejemplo del formato de ingreso automático de los términos y definiciones a la base de datos del diccionario

Cada término con sus definiciones debe ir en un archivo llamado *terminon.txt*, donde *n* corresponde a un número consecutivo en la carpeta *terminos* de la carpeta principal de SGADO, viéndose ésta como se muestra en la siguiente imagen.

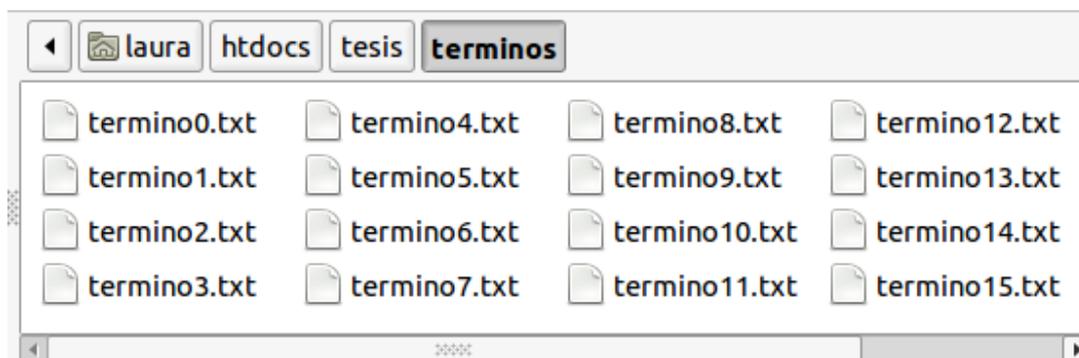


Figura 12. Carpeta contenedora de los términos con sus definiciones.

4.2.2. Paradigmas

Para tener una base de conocimientos más amplia y darle mayor flexibilidad al usuario es conveniente que se formen paradigmas con palabras con una gran relación semántica. Estos paradigmas pueden obtenerse directamente desde un diccionario de sinónimos (en este proyecto el diccionario incorporado es el Diccionario de Sinónimos Anaya), aunque pueden ingresarse otras fuentes. El único requisito consiste en que el diccionario se ingrese en un archivo de texto que cumpla con las siguientes características:

El archivo deberá contar con una codificación de caracteres tipo ISO 8859-1 con el nombre *Sinonimos.txt* ubicado dentro de la carpeta *paradigmas*. El formato de este archivo es de la siguiente manera: cada conjunto de palabras sinónimas debe ir en un renglón distinto, mientras que cada término correspondiente al conjunto deberá ir separado por un tabulador. Al final de cada línea se deberá colocar el marcador con un tabulador previo “_____”. La primera palabra de cada grupo es considerada como el índice del paradigma, para cubrir con la mejora propuesta en 3.1.1.

Por ejemplo, un archivo formado por tres conjuntos de sinónimos extraídos del diccionario de Sinónimos Anaya iría de la siguiente forma:

```

1 añejo antiguo rancio vetusto viejo arcaico remoto pretérito veterano curado _____
2 abarataarse enorgullecerse engréirse _____
3 abarcar abrazar rodear ceñir comprender englobar implicar incorporar incluir _____
  
```

Figura 13. Ejemplo del formato para el ingreso automático de los paradigmas a la base de datos del diccionario

4.2.3. Lista de paro

La lista de paro deberá encontrarse en un archivo de texto plano *stopList.txt* dentro de la carpeta *lib*. Al igual que los demás archivos, éste deberá tener una codificación de caracteres del tipo ISO 8859-1. El formato de este archivo es bastante simple, pues bastará con que se tenga un listado con cada palabra en un renglón diferente. En la siguiente imagen se puede observar una pequeña parte del archivo con la lista de paro de ejemplo.



```
10 algo
11 alguno
12 allá
13 allí
```

Figura 14. Ejemplo del formato de la lista de paro para ser ingresada automáticamente a la base de datos del diccionario

Una vez que todas las entradas cuentan con el formato aquí especificado, bastará con ejecutar el script *lematiza_base_conocimiento.sh*, para que todos los archivos sean analizados automáticamente por FreeLing y sus resultados se almacenen en nuevos archivos, listos para su ingreso automático en la base de datos.

4.3. Ingreso de las entradas a la base de datos

Se realizó un sistema en PHP que en tres pasos ingresa en el diccionario las entradas lematizadas por FreeLing y con su formato original, dependiendo de la estructura de la tabla en la que se ingresa cada uno de los datos. Además, se realizó un programa que ingresa los índices que asocian directamente los términos del diccionario con los paradigmas. Se eligió realizar el desarrollo en PHP porque se requería que la herramienta fuera en línea, además de ser de fácil uso para personas que no cuentan con conocimientos profundos de computación. Otro aspecto importante es que de esta manera se conserva la compatibilidad con el sistema de selección y depuración de resultados de Describe®, lo que permite consistencia en el trabajo del creador del diccionario.

4.3.1. Ingreso de las entradas

El primer programa, *1_inserta_definiciones.php* se encarga de insertar las definiciones en la tabla *definiciones* y los términos en la tabla *terminos_diccionario*. Durante este proceso se va mostrando una pantalla con estadísticas de los términos y el número de definiciones que se ingresó por cada uno. También ingresa las palabras individuales de las definiciones en la tabla *terminos_generales* para ir registrando el vocabulario con el que fue descrito cada término.

El segundo programa es *2_IngresalistaParo.php*. Este programa se encarga de ingresar a la tabla *lista_paro* todas las palabras vacías lematizadas. Durante la ejecución de este programa, se va desplegando un listado con las palabras que se van ingresando y al final se presenta una estadística sobre el número total de palabras ingresadas.

Finalmente, *3_Ingresaparadigmas.php* es el programa encargado de ingresar en la tabla *paradigmas* dentro de la base de datos los paradigmas provenientes del diccionario de sinónimos. Este proceso puede llegar a tardar un poco, dependiendo del tamaño del diccionario. Durante su ejecución se van mostrando estadísticas sobre el avance en el ingreso de los paradigmas. Además de ingresar los paradigmas, este sistema también agrega los vocablos que lo componen a la tabla *terminos_generales*, siempre y cuando éstos no hayan sido agregados con anterioridad por *1_inserta_definiciones.php*.

4.3.2. Ingreso de los índices

Se desarrolló el programa *4_AsocterminosDiccionario_Paradigmas.php* encargado de realizar la indización que asocia a los términos del diccionario con sus correspondientes paradigmas. Cabe señalar que para aquellas palabras que no cuentan con un paradigma particular, es decir, que no tuvieron algún sinónimo especificado durante el proceso de ingreso de paradigmas, se les crea un paradigma con único elemento en su interior. De esta manera se evita tener que consultar otras tablas para términos sin relaciones sinonímicas.

Para realizar esta indización, simplemente se busca en la tabla *definiciones* —sin considerar palabras vacías— aquellos vocablos que conformaron las definiciones de un término dado. A

continuación se procede a localizar los índices de estos vocablos en la tabla *terminos_generales* y se identifica, dentro de la tabla *paradigmas*, cuáles son los identificadores de los paradigmas que contienen dichos índices. Finalmente se ingresan en la tabla *terminos_diccionario_paradigma* el identificador del término del diccionario, el identificador del paradigma al que pertenece cada vocablo de la definición y la posición en la que dicho paradigma se encuentra dentro de la definición.

Este programa tiene una alta carga de operaciones computacionales, pues debe revisar la tabla *paradigmas* una vez por cada vocablo en las definiciones para encontrar su asociación y registrarla, por lo que su complejidad es $O(nm)$, donde n corresponde al número de términos total en las definiciones de la base de datos y m corresponde al número de términos en la tabla *paradigmas*.

El tiempo inicial de ejecución de este programa era de aproximadamente 3 horas para cerca de 1000 definiciones y 330 términos del diccionario en una computadora personal con un procesador Intel Core i7 y 6 GB en RAM. Se logró mejorar el tiempo de ejecución mediante la utilización de tablas hash en PHP y de la optimización de las búsquedas reduciendo el número de consultas a la base de datos descargando la misma a memoria local y distribuyéndola en las tablas hash. Este proceso requiere mayor uso de memoria local, por lo que las variables de solicitud de memoria para este programa tuvieron que ser modificadas, pero se pudo observar una mejora considerable en el tiempo de ejecución, pues en su versión final el programa tarda poco menos de 20 minutos en ingresar las mismas definiciones y términos. Para orientar al desarrollador del diccionario sobre el avance en el ingreso de los datos, una serie de estadísticas se pueden ver durante la ejecución de este programa.

Capítulo 5. Diccionario Onomasiológico de las Sexualidades en México (DOSEX-MX)

Para esta tesis se decidió probar la metodología propuesta en el capítulo 4, así como la mejora al motor de búsqueda del diccionario onomasiológico mediante la creación de un diccionario onomasiológico de especialidad partiendo desde cero, es decir, desde la selección de los términos del diccionario y de los paradigmas, la búsqueda de definiciones y la creación de la lista de paro. El área temática de especialidad de este diccionario es la sexualidad, que es un área que genera mucho interés por parte del sector salud y del público en general. En este capítulo se describe paso a paso el proceso que se llevó a cabo para la realización del Diccionario Onomasiológico de las Sexualidades en México (DOSEX-MX).

5.1. La base de conocimientos

La base de conocimientos que se recopiló para hacer las pruebas de generación semi-automática y de uso del DOSEX-MX constó de un total de 332 términos, 975 definiciones, 166 palabras en la lista de paro y más de 25,000 paradigmas que agrupan a cerca de 33,000 vocablos. A continuación se describen los elementos que constituyen este diccionario.

5.1.1. Términos

En el GIL actualmente se cuenta con el Corpus de las Sexualidades en México, que es un compendio de textos relativos a la sexualidad emitidos primordialmente por hablantes mexicanos. Estos textos comprenden información emitida tanto por hablantes expertos en área de la sexualidad e instituciones reconocidas en la materia, como por expresiones e ideas expuestas en foros y chats públicos. La razón de esta diversidad reside en el hecho de que se desea tener información relativa a hablantes de diversos niveles académicos y socioculturales, pues es un área en la que estos factores influyen de una manera muy importante en el léxico y en su

utilización, además de que se desea que el diccionario sea una herramienta abierta al público en general y se adapte en a las necesidades de éstos.

Lázaro (2005) obtuvo un listado de 1285 términos de sexualidad entre simples y complejos. Este listado fue el resultado de un análisis semi-automático del Corpus de las Sexualidades en México mediante la aplicación de las herramientas de *Wordsmith* descritas en el capítulo 4.1.1. De este listado automático, Lázaro (2005) seleccionó un total de 1000 términos como parte de la terminología básica del área.

Para este trabajo se decidió tomar únicamente la tercera parte de este listado para realizar las pruebas de generación del diccionario, así como de su uso y funcionamiento. Los términos seleccionados de manera aleatoria se muestran en el Apéndice A.

5.1.2. Definiciones

Las definiciones de este sistema fueron obtenidas de tres maneras diferentes: primero se comenzó utilizando Describe® en conjunto con el Sistema para la selección y depuración de los resultados de Describe®. Mediante estas herramientas se obtuvieron de forma semi-automática algunas definiciones para varios términos. Sin embargo, durante la realización de este trabajo se tuvieron algunos problemas de sobrecalentamiento del servidor en el que Describe® se encuentra alojado, por lo que fue necesario conseguir las definiciones desde otra fuente de información disponible por parte del GIL: el Sistema de Administración de Bancos Terminológicos — SABTEF— (Sierra et al., 2004), que entre sus bases de datos cuenta con un banco amplio de definiciones para términos en sexualidad. Estas definiciones fueron obtenidas de diversas páginas de internet como foros, páginas de información, de entretenimiento o de medicina. Finalmente, para aquellos términos para los que no fueron encontradas definiciones en SABTEF se comenzó a buscar sus definiciones en forma manual en foros y páginas web públicas.

En promedio cada término cuenta con tres definiciones, aunque algunos términos únicamente cuentan con una definición, mientras que en otros hay hasta cinco. Lo anterior se debe a que algunos términos son mucho más comunes que otros y el número de fuentes disponibles para ellos es bastante alto.

5.1.3. Paradigmas

Los paradigmas fueron obtenidos en su mayoría del diccionario de sinónimos Anaya para el español general. En total este diccionario cuenta con más de 25,000 grupos de sinónimos y cerca de 33,000 palabras. Al tratarse de un diccionario para español general, hubo que agregar algunos paradigmas realizados manualmente que ayudasen al diccionario a considerar términos coloquiales comúnmente utilizados en sexualidad por hablantes mexicanos. Para aumentar el tamaño de los paradigmas, se realizaron entrevistas informales a algunos alumnos de la Maestría en Ciencia e Ingeniería de la Computación de la UNAM. Algunos de los paradigmas que se agregaron manualmente son:

- *ano, detroit, culo, asterisco, aniceto, chiquito, fundo, atras*
- *pene, falo, pito, pilin, verga, reata, polla, picha, vara, platano, chile, pepino, pirrin, mastil, prieto*
- *vulva, concha, cosita, papaya, raja*
- *senos, tetas, bubis, mamas, melones*
- *testículos, huevos, tanates, criadillas, bolas, bolainas*

5.1.4. Lista de paro

La lista de paro es la misma lista que se utilizó para el diccionario de términos lingüísticos en el GIL y cuenta con 166 palabras. El listado completo de las palabras contenidas en la lista de paro puede verse en el Apéndice B.

5.2. Diccionario resultante

Al concluir la recopilación de la información que constituye la base de conocimientos descrita anteriormente, se ejecutaron los scripts y programas de SGADO y de esta manera se generaron y llenaron de forma automática las tablas de la base de datos del diccionario de sexualidad.

El algoritmo de búsqueda del diccionario onomasiológico fue desarrollado en PHP en su versión original y en su versión mejorada, con la finalidad de realizar pruebas y comparar sus resultados.

Ambos diccionarios utilizan la misma base de conocimientos, de manera que los cambios en las salidas corresponden únicamente a los cambios realizados al motor de búsqueda.

El diccionario divide los resultados en tres secciones: *muy probable*, *probable* y *poco probable*. La distribución de los resultados en estas tres secciones se basa en el criterio de ranking explicado en el capítulo 3.1.2, por lo que los elementos de cada sección se encuentran ordenados alfabéticamente, considerándose que todos ellos tienen el mismo grado de probabilidad de ser el término buscado por el usuario. La interfaz gráfica consiste en una página plana, sin mayores adornos.

Un ejemplo de salida del DOSEX-MX se muestra a continuación.

Consulta:

Muy probable	Probable	Poco probable
0 - abortar	0 - aborto	0 - abstinencia_sexual
1 - anticonceptivo	1 - androginia	1 - adolescencia
2 - espermicida	2 - andropausia	2 - adulto mayor
3 - hombre	3 - andrógino	3 - alcanzar el orgasmo
4 - homosexual	4 - anticoncepción de emergencia	4 - alquimia de el éxtasis
5 - método de el ritmo	5 - clítoris	5 - amor
6 - preservativo	6 - concepción	6 - ampullitis
7 - píldora anticonceptiva	7 - consolador	7 - andropatía
8 - violación sexual	8 - cunnilingus	8 - andrógenos
	9 - dildo	9 - anorgasmia
	10 - embarazo	10 - anovulatorio
	(...)	(...)

Figura 15. Ejemplo de salida del DOSEX-MX

La Figura 16 muestra una salida del DOSEX-MX con el algoritmo de búsqueda original, mientras que la Figura 17 muestra una salida obtenida con el algoritmo mejorado. Como se puede observar, aunque la consulta es la misma y por tanto, la entrada coincide en el número de palabras y de paradigmas, la consideración que hace el algoritmo mejorado sobre el orden de

aparición de estos últimos permite que los términos SIDA y VIH suban en el ranking y se ubiquen dentro de las primeras posiciones.

Consulta:

Muy probable	Probable	Poco probable
1 - erotismo	1 - anorgasmia	1 - abortar
2 - filosofía de el tantra	2 - anticonceptivo	2 - aborto
3 - respuesta sexual	3 - cuarenta_y_uno	3 - abstinencia sexual
4 - sífilis	4 - desarrollo sexual	4 - acrotomofilia
	5 - erótico	5 - adolescencia
	6 - eyaculación femenina	6 - adulto mayor
	7 - eyaculación precoz	7 - afrodisíaco
	8 - gonorrea	8 - alcanzar el orgasmo
	9 - género	9 - amantes
	10 - mujer	10 - amenorrea
	11 - orgasmo	11 - amor
	<u>12 - sida</u>	12 - ampullitis
	<u>13 - vih</u>	13 - andropausia
		(...)

Figura 16. Salida del DOSEX-MX para una consulta con el algoritmo de búsqueda original

Consulta:

Muy probable	Probable	Poco probable
<u>1 - sida</u>	1 - acrotomofilia	
<u>2 - vih</u>	2 - bestialidad	
	3 - climaterio	
	4 - clítoris	
	5 - desarrollo sexual	
	6 - deseo	
	7 - deseo sexual	
	(...)	

Figura 17. Salida del DOSEX-MX para una consulta con el algoritmo de búsqueda mejorado

La siguiente ventana muestra una salida del diccionario en su versión mejorada en la que se puede apreciar la consideración de la utilización de lenguaje coloquial por parte de los usuarios gracias a los paradigmas ingresados en forma manual.

Consulta:

Muy probable	Probable	Poco probable
1 - analgia		
2 - coito anal		
3 - herpes genital		
4 - pene		
5 - penetración		
6 - violación sexual		

Figura 18. DOSEX-MX mejorado considerando lenguaje coloquial

El tiempo de respuesta de ambos algoritmos es muy similar y ambos cumplen con el objetivo de devolver una respuesta al usuario en tiempo real. En el siguiente capítulo se hablará de la evaluación del algoritmo de búsqueda mejorado en comparación con el algoritmo original.

Capítulo 6. Evaluación y resultados

Para realizar una evaluación sobre la mejora en el motor de búsqueda del diccionario onomasiológico, se analizaron diversos métodos de evaluación utilizados en sistemas similares.

6.1. Criterios de evaluación

Se encontró en el trabajo de Bilac *et al.* (2004) que el método de evaluación de su diccionario inverso consistía en tomar entradas aleatorias de diccionarios que no coincidieran con aquellos que se habían utilizado en la alimentación de la base de conocimientos de su diccionario. Este método de evaluación puede dar resultados falsamente alentadores, pues dado que la alimentación de la base de conocimientos incluye definiciones formales, éstas suelen estar escritas en un lenguaje particular y común a las definiciones del propio diccionario onomasiológico que no es necesariamente el tipo de lenguaje que un usuario no experto en el área puede manejar. Entonces, en realidad no se realiza una evaluación con el público objetivo para el cual fue creado el diccionario y, por lo tanto, este método de evaluación no se considera apropiado.

Dutoit y Nugues (2002) presentan una evaluación en su diccionario para el francés realizada mediante el ingreso de definiciones en lenguaje natural que los propios autores establecieron, por lo que la imparcialidad de esta evaluación se pone en duda, pues al conocer el modo de funcionamiento del diccionario y sus fortalezas y debilidades, es fácil que los propios autores ingresen consultas a modo para mostrar mejores resultados, de tal forma que este método de evaluación tampoco parece pertinente.

Por su parte, El-Kahlout y Oflazer (2004) muestran un método de evaluación consistente en dos versiones: la primera contempla definiciones ingresadas por usuarios que no intervinieron durante la creación del diccionario y la segunda consiste —de manera análoga a Bilac *et al.* (2004)— en considerar como entrada definiciones obtenidas a partir de otros diccionarios. La primera alternativa de estos autores fue considerada como la más objetiva, aunque los resultados de esta evaluación únicamente indican el porcentaje de veces que la palabra buscada se encontró

entre los primeros cincuenta resultados, pero no se precisa cuál es la ubicación promedio del término buscado entre las listas de términos sugeridos.

Por lo anterior, se decidió realizar la evaluación mediante definiciones obtenidas de usuarios sin relación con el sistema y considerar el posicionamiento promedio de la palabra objetivo, además de si ésta es encontrada o no dentro de la salida del DOSEX-MX. Las definiciones utilizadas para esta evaluación consistieron en definiciones extraídas de un experimento de terminología llevado a cabo en el GIL con anterioridad al desarrollo del DOSEX-MX que se describe a continuación.

6.2. Definiciones de los usuarios

En el GIL se llevó a cabo un experimento de terminología que consistía en lo siguiente:

A un grupo de estudiantes de nivel licenciatura de las carreras de Ingeniería en Computación y Licenciatura en Licenciatura en Lengua y Literaturas Hispánicas, en números iguales, se les pidió que se dividieran en dos subgrupos *A* y *B*. El subgrupo *A* recibía una lista con algunos términos, mientras que el subgrupo *B* recibía una lista con términos diferentes a los del subgrupo *A*. A continuación se les pedía a ambos subgrupos que escribieran en un papel la definición que utilizarían para hacer que una persona comprendiera cada uno de los términos que les había tocado, sin utilizar el término mismo. Una vez que todos los alumnos de ambos subgrupos habían terminado de ingresar en un papel sus definiciones, se les pidió que intercambiaran sus definiciones, sin revelar los términos que les habían tocado. Ahora el subgrupo *A* debía tratar de adivinar qué términos estaban describiendo sus compañeros del grupo *B* y viceversa y anotarlos junto a las definiciones.

Para este experimento se utilizaron algunos términos de sexualidad, por lo que se cuenta con definiciones apropiadas para la temática del diccionario. Además, otra ventaja que se tiene es que se cuenta con el intento que realiza un ser humano durante la interpretación de cada definición, por lo que se tiene un punto de referencia adicional al momento de evaluar. Los términos que los alumnos debían definir fueron: *homosexualidad*, *identidad sexual*, *herpes*, *condón*, *bisexual*, *género* y *SIDA*.

6.3. Evaluación

Se realizó un análisis ingresando al DOSEX-MX —tanto al del algoritmo mejorado como al del original— cada una de las definiciones de los usuarios, tal cual habían sido escritas por éstos y se registraron las posiciones tanto relativas —la posición dentro de una de las secciones en particular (muy *probable*, *probable* o *poco probable*)— como absolutas —la posición de la palabra (o palabras) objetivo en el ranking general.

Es importante destacar que algunas definiciones de los usuarios podían ser aplicadas a más de un término dentro del diccionario. Por ejemplo, la definición “*enfermedad de transmisión sexual*” puede hacer referencia a cualesquiera enfermedades de transmisión sexual que se encuentren entre los términos del DOSEX-MX. En estos casos se analizaron todos los términos que eran congruentes con la definición del usuario y se obtenía un promedio de su ubicación en la tabla general.

Un aspecto importante que se debe observar al momento de determinar la mejora del motor de búsqueda del diccionario onomasiológico consiste en la reducción del número de términos candidatos mostrados a la salida. Una reducción significativa en la cantidad de términos candidato representa una mejora para el usuario, que tardará menos tiempo en identificar el término que busca, siempre y cuando esta reducción de términos candidatos no excluya el término objetivo.

Debido a que cada definición arroja un número diferente de términos candidato, en lugar de interpretar los resultados por posiciones o número de términos de diferencia, éstos se interpretarán con porcentajes. Las tablas generadas por cada definición durante el proceso de experimentación se encuentran en el Apéndice C. A continuación se muestra un análisis de los resultados obtenidos a partir de dichas tablas.

6.3.1. Resultados para el término *homosexualidad*

El término *homosexualidad* tuvo una mejora promedio del 33.91% en el posicionamiento absoluto del término objetivo dentro del total de los términos de salida del algoritmo mejorado, siendo la definición 5 —*Gusto sexual por personas del mismo sexo*— la que representó el peor desempeño para este término (-62.50%), mientras que la definición 3 —*Preferencia por una persona que tiene el mismo sexo*— fue la que obtuvo mejores resultados (75%) en cuanto a posicionamiento general. En cuanto al número de términos candidatos, para todas las definiciones se tuvo una reducción correspondiente al 75.47%, que es su equivalente en mejoría, entregando en promedio únicamente 36 términos candidatos. El DOSEX-MX mejorado fue capaz de encontrar el término buscado en todos los casos, colocando el término en promedio en la séptima posición, encontrándose en cuatro ocasiones en la sección de términos “muy probable”, en una ocasión en el listado de “probable” y en una ocasión en “poco probable”.

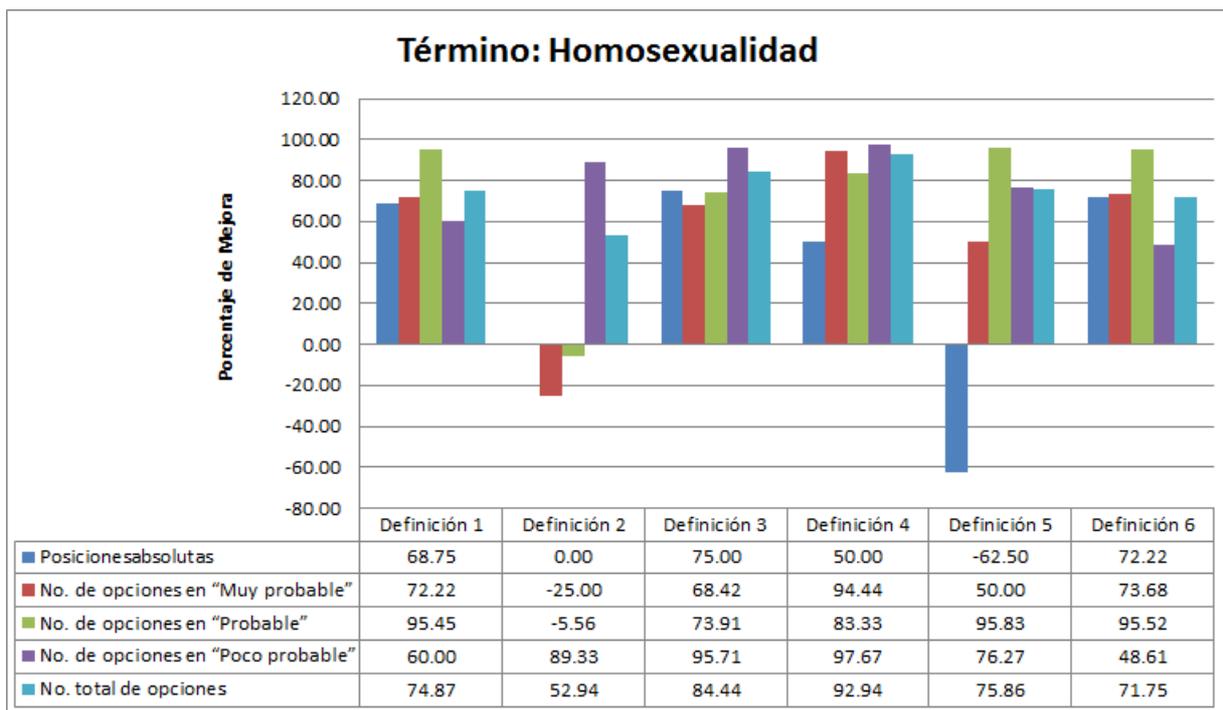


Figura 19. Gráfica de mejoras en cada rubro para el término *homosexualidad*

6.3.2. Resultados para el término *identidad sexual*

El término *identidad sexual* presentó una mejora promedio del 28.43% en cuanto al posicionamiento global del término objetivo dentro del total de los términos de salida del algoritmo mejorado, siendo incapaz de encontrar el término con las definiciones 8 —*Determinación o inclinación hacia un sexo determinado*—, 10 —*Conducta que determina el gusto e inclinación hacia un género sexual*— y 12 —*Orientación sexual de un individuo en relación a sus prácticas sexuales*—. Para la definición 10 el humano tampoco fue capaz de adivinar el término objetivo dada la descripción.

En cuanto al número de términos candidatos, para todas las definiciones se tuvo una reducción promedio correspondiente al 76.37%, que equivale a su porcentaje en mejoría, entregando en promedio únicamente 31 términos candidatos. El DOSEX-MX colocó el término en promedio en la vigésima posición, encontrándose en tres ocasiones en la sección de términos “probable” y en una ocasión en “poco probable”.

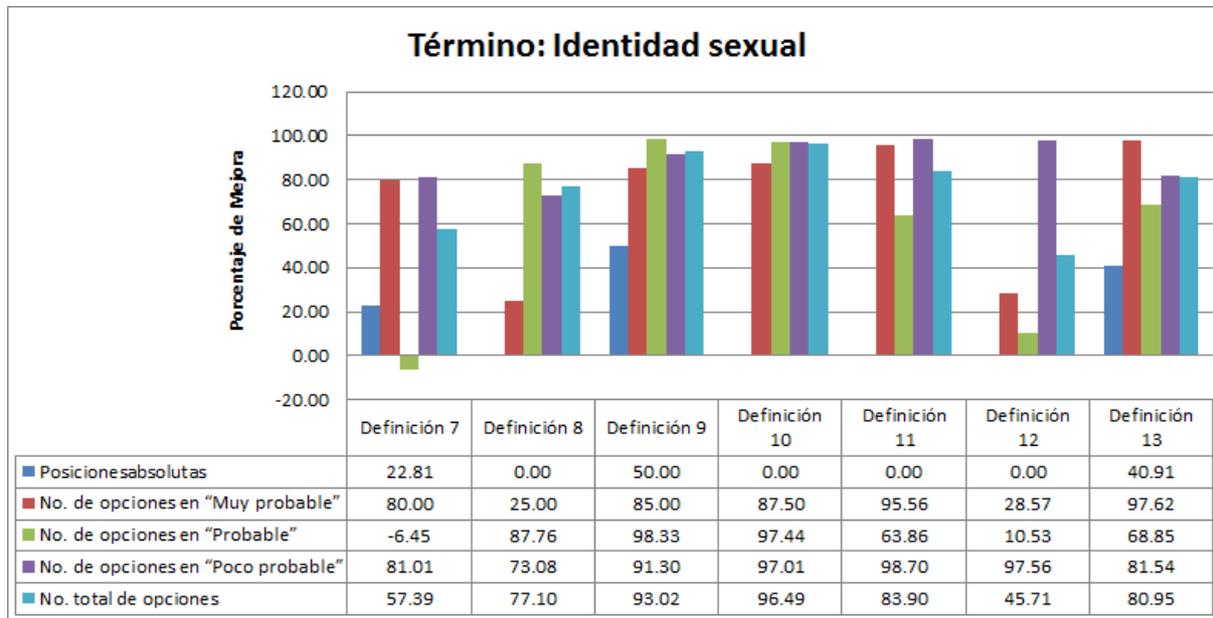


Figura 20. Gráfica de mejoras en cada rubro para el término *identidad sexual*

6.3.3. Resultados para el término *herpes*

Para el término *herpes* el DOSEX-MX con el algoritmo de búsqueda mejorado presentó una mejora promedio del 36.81% en cuanto al posicionamiento global del término objetivo dentro del total de los términos de salida, encontrando en todos los casos una respuesta apropiada para cada definición proporcionada. En este caso muchas de las definiciones carecían de la especificidad requerida para que el DOSEX-MX discriminara de entre sus opciones. Por ejemplo, las definiciones 15 —*Enfermedad transmitida por el contacto sexual*— y 18 —*Enfermedad de transmisión sexual*— pueden hacer referencia a una gran variedad de términos y no exclusivamente a *herpes*. De hecho, para este término ningún humano fue capaz de identificar el término objetivo con las descripciones proporcionadas.

En cuanto al número de términos candidatos, para todas las definiciones se tuvo una reducción promedio correspondiente al 62.75%, entregando en promedio 73 términos candidatos. Esta gran cantidad de datos se debió, como se mencionó en el párrafo anterior, a la falta de detalles por parte de quien proporcionaba la definición. El DOSEX-MX colocó el término en promedio en la trigésima posición, encontrándose en todas las ocasiones en la sección de términos “muy probable”.

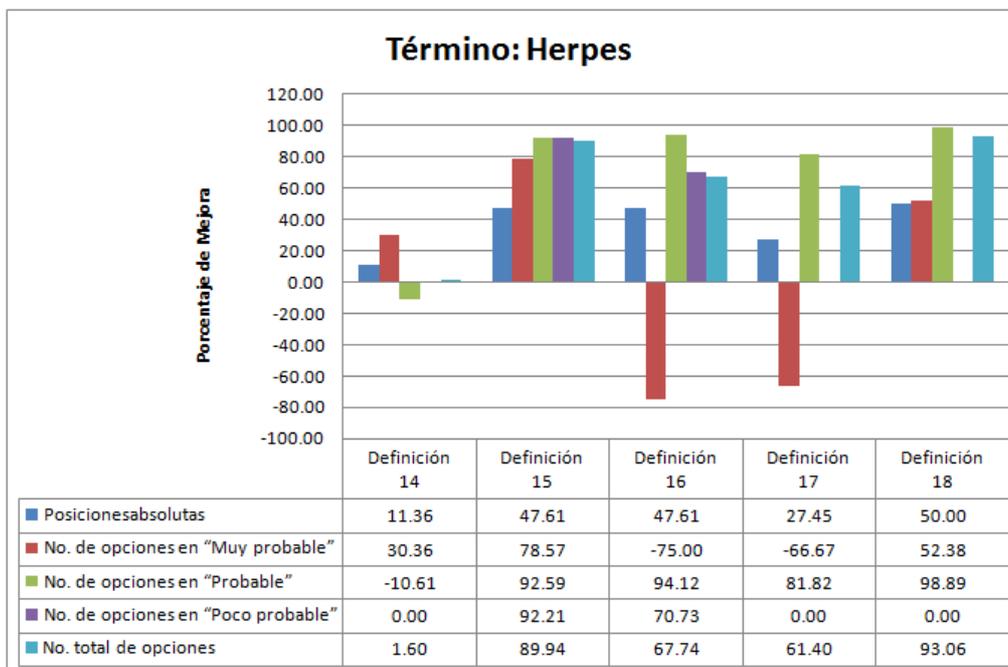


Figura 21. Gráfica de mejoras en cada rubro para el término *herpes*

6.3.4. Resultados para el término *condón*

El término *condón* con el algoritmo de búsqueda mejorado presentó una pérdida del 21.49% en cuanto al posicionamiento global del término objetivo dentro del total de los términos de salida. El diccionario no fue capaz de identificar el término con la definición 20 —*Método anticonceptivo fácil de usar y de adquirir*—, pues esta definición no se adapta a las definiciones de *condón* en la base de conocimientos, ya que en estas se enfocan en el material del que se encuentran hecho, su forma y su modo de uso, sin hacer referencia a que se trata de un método anticonceptivo o a que es de fácil adquisición.

En cuanto al número de términos candidatos, para todas las definiciones se tuvo una reducción promedio correspondiente al 70.74%, entregando en promedio únicamente 16 términos candidatos. El DOSEX-MX colocó el término en promedio en la decimosegunda posición, encontrándose en 3 ocasiones en la sección de términos “muy probable” y una única ocasión en la sección de “probable”.

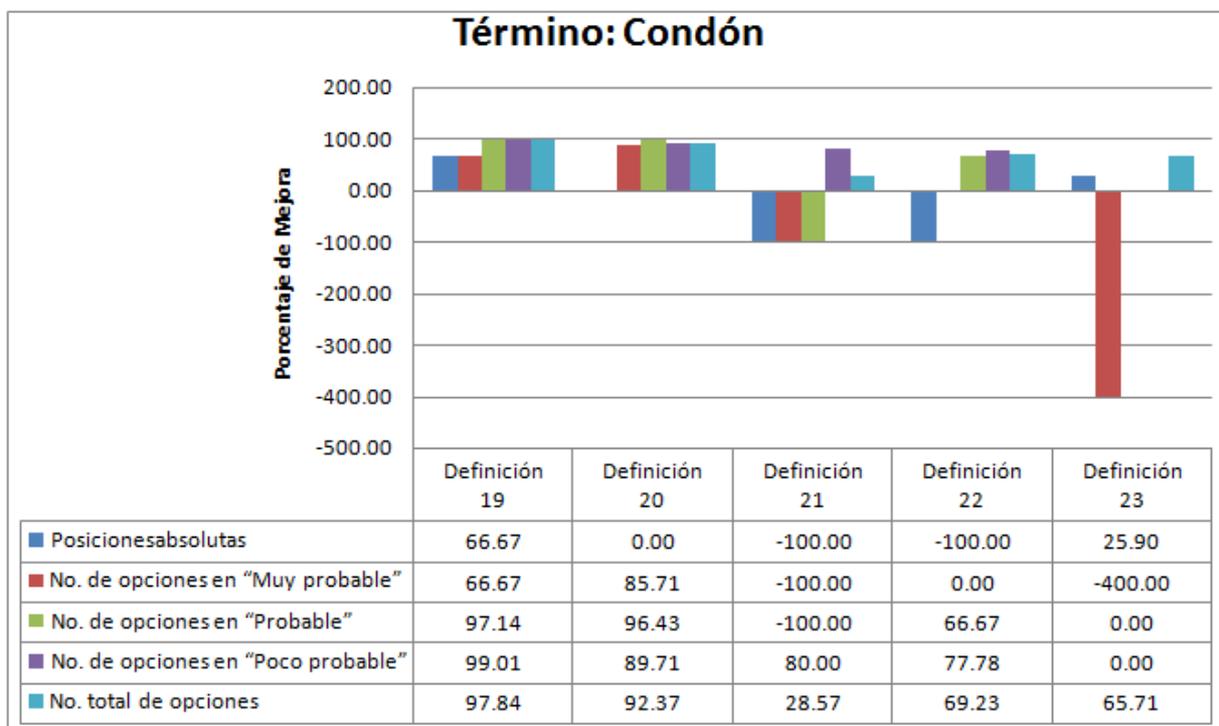


Figura 22. Gráfica de mejoras en cada rubro para el término *condón*

6.3.5. Resultados para el término *bisexual*

Para el algoritmo de búsqueda mejorado el término *bisexual* presentó una mejora promedio del 8.33% en cuanto al posicionamiento global del término objetivo dentro del total de los términos de salida. El diccionario no fue capaz de identificar el término en tres ocasiones, pues sólo cuenta con única definición para *bisexual* en su base de conocimientos —*Atracción sexual por personas de ambos sexos y / o que tiene relaciones sexuales indistintamente con ellas*—.

En cuanto al número de términos candidatos, para todas las definiciones se tuvo una reducción promedio correspondiente al 74.10%, entregando en promedio 37 términos candidatos. El DOSEX-MX colocó el término en promedio en la primera posición, encontrándose, por lo mismo, en las 3 ocasiones en la sección de términos “muy probable”.

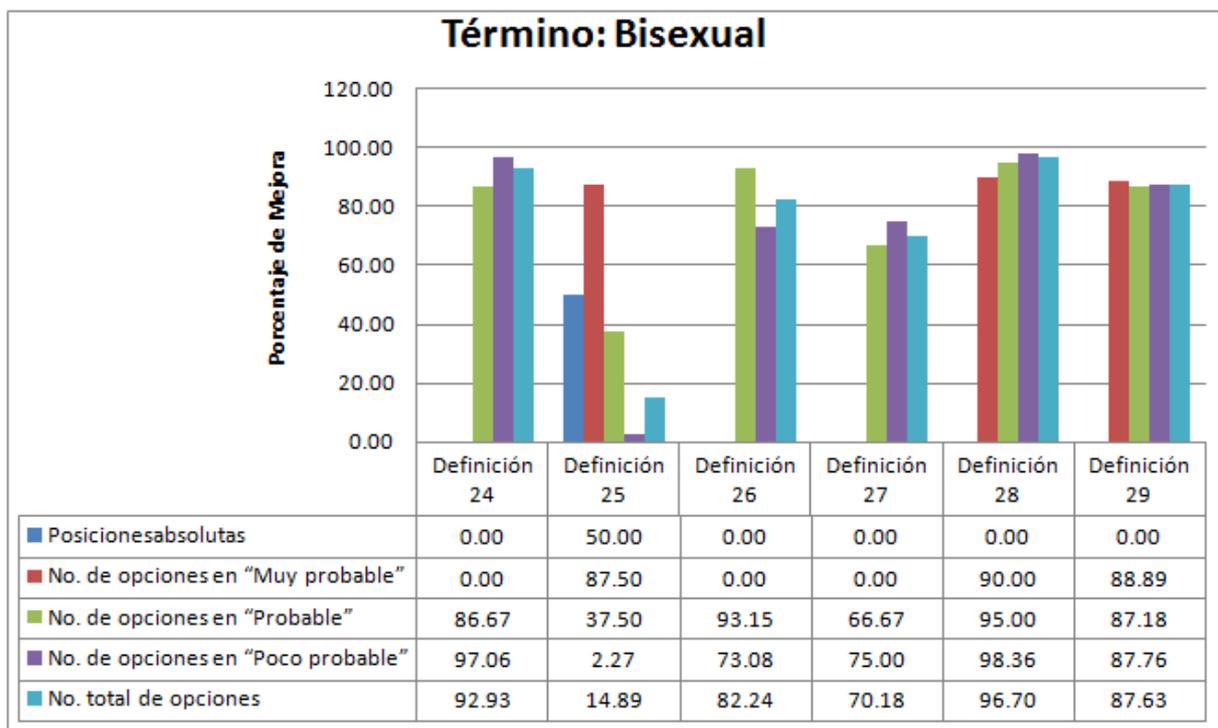


Figura 23. Gráfica de mejoras en cada rubro para el término *bisexual*

6.3.6. Resultados para el término género

El término *género* tuvo una mejora promedio del 25.00% en el posicionamiento absoluto del término objetivo dentro del total de los términos de salida del algoritmo mejorado, aunque el DOSEX-MX mejorado fue incapaz de encontrar el término buscado en dos ocasiones. El término *género* resulta un poco conflictivo para definir por las personas e incluso éstas no pudieron adivinar el término en una ocasión. El DOSEX-MX con el algoritmo antiguo identificó el término en las tres ocasiones, pero entregaba en promedio al usuario un total de 110 términos candidatos, por lo que, aunque sí fue capaz de entregar una respuesta correcta, el usuario tenía que buscar entre listados muy grandes el término que deseaba encontrar.

En cuanto al número de términos candidatos, para todas las definiciones se tuvo una reducción correspondiente al 90.75%, que es su equivalente en mejoría, entregando en promedio únicamente 6 términos candidatos. El DOSEX-MX colocó el término objetivo en tercera posición, encontrándose en la sección de términos “probable”. Éste es el término para el que existió una mayor reducción en el número de términos candidatos.

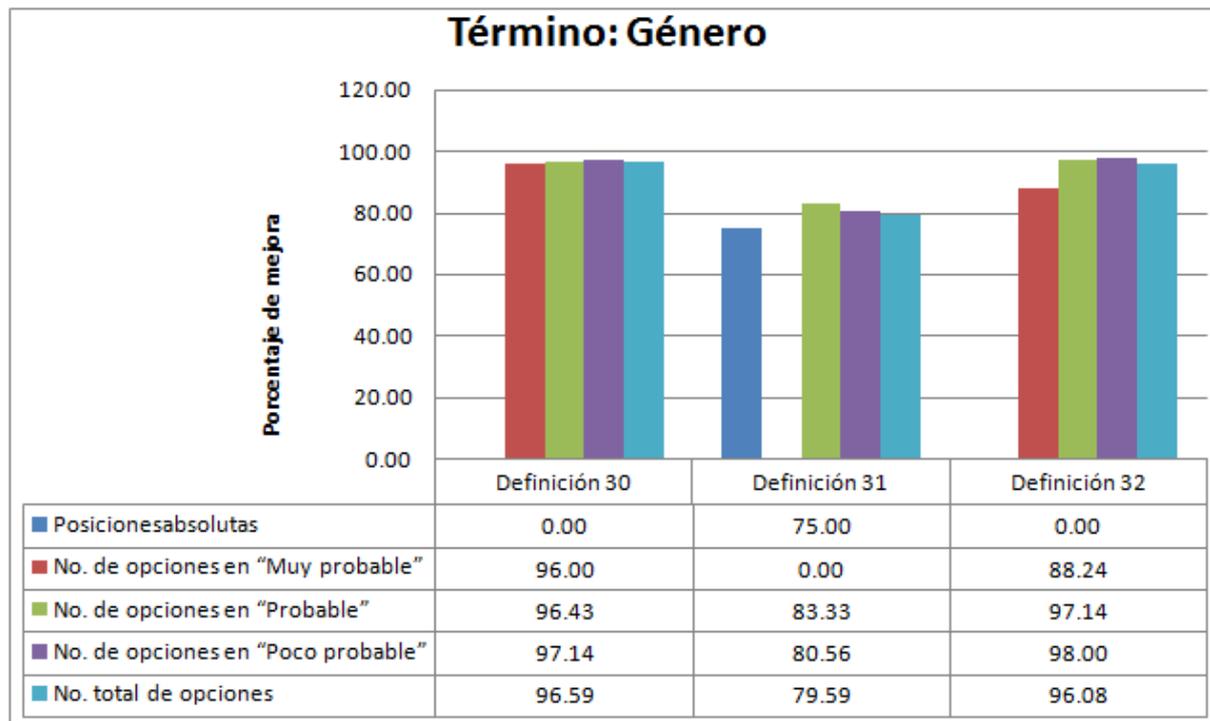


Figura 24. Gráfica de mejoras en cada rubro para el término *género*

6.3.7. Resultados para el término *SIDA*

El término *SIDA* presentó una disminución del 5.00% en cuanto al posicionamiento global del término objetivo dentro del total de los términos de salida del algoritmo mejorado, siendo incapaz de encontrar el término en dos ocasiones. Las definiciones para las que no pudo dar una respuesta satisfactoria fueron la 33 —*Enfermedad donde las defensas del cuerpo son muy bajas. Se contagia por medio de relaciones sexuales*— y la 38 —*Enfermedad de transmisión sexual incurable y altamente contagiosa*—. El algoritmo original tampoco fue capaz de identificar al término *SIDA* como respuesta debido a que en la base de conocimientos sólo existe una única definición para este término: *Síndrome de Inmuno Deficiencia Adquirida. Enfermedad infecciosa, crónica y mortal que destruye las defensas del organismo exponiendo lo a múltiples infecciones que pueden ser desde leves hasta muy graves.*

En cuanto al número de términos candidatos, para todas las definiciones se tuvo una reducción promedio correspondiente al 78.53%, que equivale a su porcentaje en mejoría, entregando en promedio únicamente 23 términos candidatos. El DOSEX-MX colocó el término en promedio en la cuarta posición, encontrándose en dos ocasiones en la sección de términos “muy probable” y en una ocasión en “probable”.

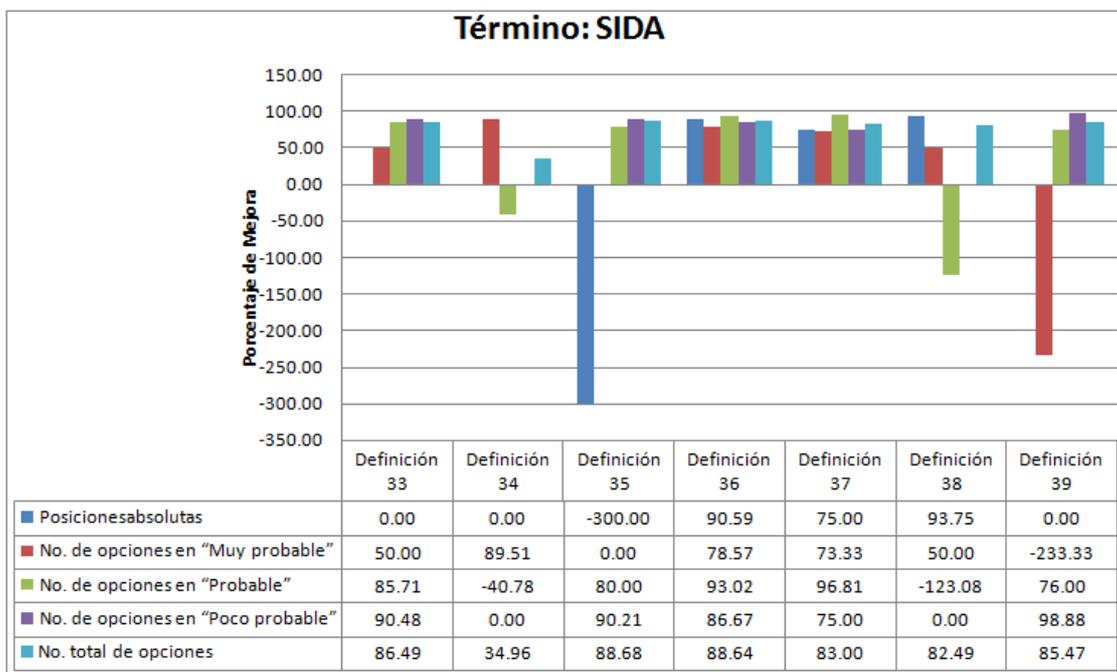


Figura 25. Gráfica de mejoras en cada rubro para el término *SIDA*

6.3.8. Evaluación general

En términos generales se obtuvo en promedio una mejora en el posicionamiento del término objetivo en la tabla general del 15% con respecto a los resultados obtenidos con el algoritmo de búsqueda original. Esto representa que en cada salida el término objetivo se colocó en promedio un 15% más arriba que los demás términos, lo que agiliza considerablemente su localización por parte del usuario.

También se puede observar que se tuvo una mejoría dramática en el porcentaje de reducción del número de términos candidatos, pues en promedio se redujo en 75.53%, lo que permite que el usuario busque el término entre listados más cortos, mejorando de esta forma la calidad de los resultados. Sin embargo, esta mejora resultó también en una reducción en el número de veces que el diccionario encontró la respuesta apropiada, pues mientras el diccionario con el algoritmo original pudo encontrar la respuesta el 97% de las ocasiones, la mejora propuesta sólo obtuvo una precisión del 71%. De aquí se concluye que el porcentaje de efectividad del diccionario disminuyó, pero la calidad de sus resultados aumentó.

Al compararlo con los resultados de El-Kahlout y Oflazer (2004), que utilizaron un método de evaluación similar, se puede ver que el 71% de buenos resultados obtenidos por DOSEX-MX queda en un nivel superior respecto al 66% de efectividad obtenido por ambos autores. Esta efectividad se mide haciendo referencia a si el sistema es capaz de identificar el término que el usuario busca listándolo entre los primeros cincuenta términos resultantes.

Es importante resaltar que los términos de sexualidad, como *identidad sexual*, *género* o *herpes* son términos que a los propios humanos les cuesta trabajo expresar o de los que desconocen, por lo que sus descripciones tienden a ser confusas y vagas. Estos son factores que influyen también al momento de llevar a cabo la evaluación, pues sus resultados dependen en gran medida del nivel cultural y social de sus participantes.

Conclusiones

En esta tesis se presentó una metodología para crear de forma semi-automática diccionarios onomasiológicos especializados cuyos motores de búsqueda se enfocan a la búsqueda mediante el uso de paradigmas. Esta metodología fue utilizada con resultados satisfactorios para crear el Diccionario Onomasiológico de las Sexualidades en México. Partiendo de esta experiencia se puede ver que la metodología aquí propuesta es factible y eficaz para realizar la tarea para la que fue creada, aunque no se han realizado pruebas formales de su funcionamiento por parte de personas ajenas al proyecto.

También se dio a conocer en esta tesis una propuesta de mejora al motor de búsqueda del diccionario onomasiológico original del Grupo de Ingeniería Lingüística. La evaluación de los resultados obtenidos en esta propuesta muestra que se consiguió una mejora promedio del 15% en la mejor ubicación del término objetivo buscado por el usuario. Además, se logró una reducción promedio de casi el 76% de los listados de términos candidatos en la salida, lo que permite al usuario buscar en listas más cortas y más exactas. Esta mejora, sin embargo, redujo al 71% de efectividad del diccionario, pues en algunos casos el término objetivo era descartado por el proceso de reducción de la lista de términos candidatos.

Al realizar una comparativa entre los resultados obtenidos con el motor de búsqueda aquí propuesto y los resultados del diccionario onomasiológico del diccionario onomasiológico para el idioma turco propuesto por El-Kahlout y Oflazer (2004), se puede ver que el diccionario con el nuevo motor de búsqueda presenta una efectividad 5% superior.

Aún falta mucho por mejorar en los resultados del diccionario onomasiológico y existen actualmente diversos grupos de investigación alrededor del mundo que, como Bilac, Dutoit, Nugues, El-Kahlout, Oflazer y Zock, intentan encontrar la manera de incrementar el desempeño de este tipo de diccionarios. Los resultados obtenidos en este trabajo son alentadores, pues brindan una pauta para mejorar, pero son todavía altamente perfectibles, siendo la aplicación de métodos estadísticos una opción que parece viable.

Bibliografía

1. Alarcón, R. y Sierra, G. (2003). *The role of verbal predications for definitional context extraction*. Terminologie et Intelligence Artificielle (TIA 2003). Université de Strasbourg. Strasbourg, Francia (págs. 11-20).
2. Alarcón, R., Bach, C., y Sierra, G. (2008). *Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica*. Revista Española de Lingüística 37, pp. 247-278.
3. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L. y Padró, M. (2006). *FreeLing 1.3: Syntactic and semantic services in an open-source NLP library*. Proceedings of the fifth international conference on Language Resources and Evaluation, ELRA. Genoa, Italy.
4. Baldinger, K. (1970). *Teoría semántica: Hacia una semántica moderna*. Ediciones Alcalá.
5. Bilac, S., Watanabe, W., Hashimoto, T., Tokunaga, T. y Tanaka, H. (2004). *Dictionary search based on the target word description*. Proceedings of the Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004) (págs. 556-559).
6. Bouzy, C. (1992). *El Tesoro de la Lengua Castellana o Española: Sebastián Covarrubias en el laberinto emblemático de la definición*. Universidad de Metz. Criticón No. 54. (págs 127-144).
7. Buendía, N., Delgado, M., Márquez, C. y Méndez, J. (2005). *Diseño de un Diccionario Onomasiológico para la Web*. Facultad de Ingeniería, UNAM / Facultad de Filosofía y Letras, UNAM (tesis de licenciatura).
8. Cabrera, L. (2011). *TF-IDF Para la Obtención Automática de Términos y su Validación Mediante Wikipedia*. Facultad de Ingeniería, UNAM / Facultad de Filosofía y Letras, UNAM (tesis de licenciatura).
9. Carreras, X., Chao, I., Padró, L. y Padró, M. (2004). *FreeLing: An Open-Source Suite of Language Analyzers*. Proceedings of the 4th International Conference on Language Resources and Evaluation.

10. Dutoit, D. y Nugues, P. (2002). *A Lexical Database and an Algorithm to Find Words from Definitions*. En Frank van Harmelen (ed.): ECAI2002, Proceedings of the 15th European Conference on Artificial Intelligence, Lyon (págs. 450-454).
11. El-Kahlout I. D. y K. Oflazer (2004). *Use of Wordnet for Retrieving Words from Their Meanings*. 2nd Global WordNet Conference, Brno.
12. Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
13. Gelbukh, A. y Sidorov, G. (2006). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. Instituto Politécnico Nacional. México, D.F.
14. Lázaro, J. (2005). *Extracción de la terminología básica de las sexualidades en México a partir de un corpus lingüístico*. Facultad de Filosofía y Letras, UNAM (tesis de licenciatura).
15. Malkiel, Yakov (1975). *A typological classification of dictionaries on the basis of distinctive features*. En F.W. Housholder y S. Saporta (Eds.): *Problems in Lexicography*, (Indiana University).
16. Miller, George A. (1995). *WordNet: A Lexical Database for English*. *Communications of the ACM* Vol. 38, No. 11: 39-41.
17. Padró, L., Collado, M., Reese, S., Lloberes, M. y Castellón, I. (2010). *FreeLing 2.1: Five Years of Open-Source Language Processing Tools*. Proceedings of 7th Language Resources and Evaluation Conference, ELRA La Valletta, Malta.
18. San Segundo, R. (1996). *Sistemas de organización del conocimiento. La organización del conocimiento en la bibliotecas españolas*. Madrid: Universidad Carlos III de Madrid.
19. Shcherba, L.V. (1995). *Towards a general theory of lexicography*. *International Journal of Lexicography*. **Reimpreso de su versión original, 1940** (págs 314-350).
20. Sierra, G. (1999). *Design of a concept-oriented tool for terminology*. UMIST, Manchester (tesis de doctorado).
21. Sierra, G., Garduño, G., Moreno, M. y Castillo, G. (2004). *Banco terminológico, una base de conocimientos léxica multipropósito*. En Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués, IX Congreso Iberoamericano de Inteligencia Artificial (IBERAMIA), Tonantzintla, Puebla.

22. Sierra, G., y McNaught, J. (2000). *Design of an onomasiological search system: A concept-oriented tool for terminology*. COGALEX '08, (págs. 32 - 38).
23. Sierra, Gerardo (2000). *The onomasiological dictionary: a gap in lexicography*. Proceedings of Euralex-2000. Stuttgart, Alemania. (págs. 223-235).
24. Sierra, Gerardo (2007). *Lexicografía onomasiológica, un modelo constructivo*. V Congreso de Investigaciones Lingüístico-Filológicas, Academia Peruana de la Lengua.
25. Sierra, G. y Hernández, L. (2012). *A Proposal for Building the Knowledge Base of Onomasiological Dictionaries*. Cognitive Science. (En prensa).
26. Svensén, Bo (1993). *Practical lexicography: principles and methods of dictionary-making*. Oxford: Oxford University Press.
27. Vossen, Piek (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic*. Computers and the Humanities. Springer. Volúmen 32, Números 2-3.

Apéndice A. Términos del DOSEX-MX

abortar	bestialidad	contagio de enfermedades
aborto	bestialismo	continencia
abstinencia	bisexual	contracepción hormonal
abstinencia sexual	bisexualidad	control
acrotomofilia	blenorragia	control natural de la natalidad
actividad sexual	blenorrea	coprofilia
acto sexual	buga	cópula
adolescencia	cápsula cervical	copular
adolescente	capuchón cervical	crema espermicida
adulto mayor	características sexuales secundarias	crisis
afrodisíaco	castración	crisis de pareja
agalmatofilia	celibato	cromosomas
alcanzar el orgasmo	cervix	cuarenta y uno
alquimia del éxtasis	chancro	cuerpo
amantes	chlamydia	cuerpos cavernosos
amenorrea	ciberpornografía	cunnilingus
amor	cibersexo	cunninlingus
ampullitis	ciclo menstrual	curiosidad intelectual
analgia	circuncisión	derecho a el placer
andrógenos	cistitis	desarrollo sexual
androgenia	clamidia	deseo
andrógino	climaterio	deseo sexual
andromimetofilia	clímax	deseo y placer
andropatía	clítoris	desequilibrios de la identidad genérica
andropausia	coito	desfloración
anhedonia eyaculatoria	coito anal	detumescencia
ano	coito interrumpido	diafragma
anorgasmia	coito reservatus	diana cazadora
anorquidia	coito y orgasmo	dildo
anovulatorio	coitus interruptus	disforia genérica
anticoncepción de emergencia	colposcopia	disfunción
anticonceptivo	comida afrodisiaca	disfunción eréctil
areola	comportamiento sexual	disfunción sexual
asexual	concepción	dismenorrea
atocia	condón	dispaurenia
autoerótico	condón femenino	dispositivo intrauterino
autoerotismo	conducta sexual	dolor
autoestima	conducto deferente	drogas y sexualidad
autoestimulación	consolador	ducha
bacteria	contacto	ducha genital

educación	glándulas endocrinas	obtención del placer
educación sexual	gonorrea	ogino
embarazo	hermafrodita	oral
embarazo no deseado	herpes genital	orgasmo
emoción	heterosexual	orgasmo seco
episiotomía	hijos	orientación sexual
erección	hipertrofia	ovarios
erótico	hombre	padres
erotismo	homosexual	parafilia
esmegma	hormonas	pareja
esperma	hormonas sexuales	pasión
espermicida	identidad genérica	pedofilia
espermicidas vaginales	identidad sexual	pelonas
esponja espermicida	impotencia	pelvis
estereotipo	impulso sexual	pene
esteroides	incesto	penetración
estimulación	información	período de fertilidad
estimulación sexual	inhibición del deseo	persona
estímulo	inquisición	perversiones sexuales
estrógeno	instinto sexual	píldora
etapa	intimidad	píldora anticonceptiva
ets	levonova	placer
excitación	machismo	placer sexual
exhibicionismo	machismo invisible	precoz
excitación sexual	mamas	preservativo
eyaculación	masculino	prevención de enfermedades
eyaculación femenina	masoquismo	problemas sexuales
eyaculación precoz	masturbación	progestinas
familia	matrimonio	prohibición
femenino	menopausia	Prostitución
fetichismo	método del ritmo	psicodinámica de las parafilias
feto	método de la barrera	Pubertad
filosofía del tantra	métodos naturales	Pudor
género	miógrafo vaginal	punto g
genitales	moral	Realidad
genitalidad	mujer	Regla
gerontofilia	mujer menopáusica	Relación
glándula hipófisis	necesidad	relación sexual
glándulas	necrofilia	Represión
glándulas de cowper	ninfomanía	represión sexual

reproducción	tiempo refractario interorgásmico
reproducción asistida	tipificación sexual
respuesta sexual	tomás de aquino
rol	trabajadora sexual
rol genérico	transexual
roles sexuales	transmisión sexual
sádico	trastorno
sadismo	trastorno del deseo
sadomasoquismo	trasvestismo
salpinguitis	trauma
salud	travesti
san agustín	tricomoniasis
satisfacción	trieja
satisfacción sexual	triografía
secreción	trompas de falopio
semen	unión sexual
sensación	uretra
sensación de satisfacción	uretritis
sensualidad	urofilia
sentimiento	útero
sexo	vagina
sexo oral	vaginal
sexología	vaginitis
sexual	vasectomía
sexualidad	vello púbico
sexualidad humana	verrugas genitales
sida	vesícula seminal
sífilis	viagra
sistema de punto sexual fantasma	vibrador
sociedad	vibradores
sociosexual	vida sexual
solicitud	vih
status	vínculo
tamaño de genitales	violación
teatralización	violación sexual
técnica para eyaculación precoz	virus
temperatura basal del cuerpo	voyerismo
terapia de reemplazo hormonal	vulva
testículos	zonas erógenas
testosterona	zoofilia

Apéndice B. Lista de paro del DOSEX-MX

a	cuan	fuera	pus	usted
aca	cuando	fueron	que	vosotros
acercar	cuanto	haber	quien	ustedes
acerca	cuyo	ha	quienes	va
actualmente	cuya	hacer	quiza	varios
ademas	de	hasta	quizas	varias
ahi	demas	ir	se	vez
ahora	dentro	junto	seguir	y
ai	desde	la	seguido	ya
el	despues	las	segun	yo
algo	dicha	le	si	
alguno	decir	les	sigue	
alguna	dicho	lo	siguiente	
algunos	dar	los	siguientes	
algunas	dio	luego	sino	
alla	donde	mas	siquiera	
alli	e	me	so	
alto	ella	mi	sobre	
ante	ellos	mientras	solo	
antes	ellas	miar	su	
apenas	ello	mio	sus	
aquel	embargar	mis	suyo	
aqui	embargo	nada	tal	
asi	en	nos	tambien	
aun	entonces	nosotros	tampoco	
aunque	entre	nuestro	tan	
bajo	era	o	tanto	
cada	ser	onde	te	
casi	eran	ora	tener	
cerca	es	orita	ti	
cierto	ese	oscuro	todavia	
como	esa	pa	todo	
comun	esas	para	tras	
con	eso	pero	tu	
contra	esos	por	tus	
cual	este	porque	tuyo	
cuales	esta	porques	u	
cualquiera	estar	pos	uno	
cualquier	fue	pues	uste	

Apéndice C. Tablas de experimentación y evaluación del DOSEX-MX

1) *“Preferencia de un sujeto por individuos del mismo sexo.”*

Respuesta Humana: Homosexualidad

Respuesta correcta: Homosexualidad

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	1. homosexual	MP - 16	16	36	66	85	187
Sistema mejorado:	1. homosexual	MP - 5	5	10	3	34	47
		Mejora:	68.75%	72.22%	95.45%	60.00%	74.87%

2) *“Preferencia hacia el mismo sexo”*

Respuesta Humana: Homosexual

Respuesta correcta: Homosexualidad

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	1. homosexual	MP - 5	5	8	36	75	119
Sistema mejorado:	2. homosexual	MP - 5	5	10	38	8	56
		Mejora:	0.00%	-25.00%	-5.56%	89.33%	52.94%

3) *“Preferencia por una persona que tiene el mismo sexo.”*

Respuesta Humana: Homosexualidad

Respuesta correcta: Homosexualidad

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	3. homosexual	MP - 12	12	19	46	70	135
Sistema mejorado:	1. homosexual	MP - 3	3	6	12	3	21
		Mejora:	75.00%	68.42%	73.91%	95.71%	84.44%

4) “Género sexual donde el individuo tiene preferencia por algún individuo del mismo sexo.”

Respuesta Humana: Homosexualidad

Respuesta correcta: Homosexualidad

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	4. homosexual	MP - 6	6	18	24	43	85
Sistema mejorado:	1. homosexual	P - 2	3	1	4	1	6
		Mejora:	50.00%	94.44%	83.33%	97.67%	92.94%

5) “Gusto sexual por personas del mismo sexo.”

Respuesta Humana: Homosexualidad

Respuesta correcta: Homosexualidad

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	5. homosexual	MP - 16	16	38	48	59	145
Sistema mejorado:	1. homosexual	PP - 5	26	19	2	14	35
		Mejora:	-62.50%	50.00%	95.83%	76.27%	75.86%

6) “Preferencia por personas del mismo sexo.”

Respuesta Humana: Homosexualidad

Respuesta correcta: Homosexualidad

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	6. homosexual	MP - 18	18	38	67	72	177
Sistema mejorado:	1. homosexual	MP - 5	5	10	3	37	50
		Mejora:	72.22%	73.68%	95.52%	48.61%	71.75%

7) “Gusto carnal hacia otro individuo.”

Respuesta Humana: Atracción

Respuesta correcta: Identidad sexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	7. deseo sexual 8. deseo	PP - 21 PP - 20	57 56	5	31	79	115
Sistema mejorado:	9. deseo sexual 10. deseo	PP - 10 PP - 9	44 43	1	33	15	49
Mejora:			22.81%	80.00%	-6.45%	81.01%	57.39%

8) “Determinación o inclinación hacia un sexo determinado.”

Respuesta Humana: Tendencia sexual

Respuesta correcta: Identidad sexual

	Mejores respuestas:	Posiciones Relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	11. identidad sexual	P - 21	25	4	49	78	131
Sistema mejorado:				3	6	21	30
Mejora:			N. E.	25.00%	87.76%	73.08%	77.10%

9) “Diversidad de gustos que existen en la sexualidad.”

Respuesta Humana: Bisexualidad

Respuesta correcta: Identidad sexual

	Mejores respuestas:	Posiciones Relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	12. identidad sexual	MP - 8	8	20	60	92	172
Sistema mejorado:	13. identidad sexual	P - 1	4	3	1	8	12
Mejora:			50.00%	85.00%	98.33%	91.30%	93.02%

10) “*Conducta que determina el gusto e inclinación hacia un género sexual.*”

Respuesta Humana: Atracción

Respuesta correcta: Identidad sexual

	Mejores respuestas:	Posiciones Relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones	
Sistema antiguo:	14. identidad sexual	P - 14	22	8	39	67	114	
Sistema mejorado:				1	1	2	4	
Mejora:				N. E.	87.50%	97.44%	97.01%	96.49%

11) “*Determinación de gusto sexual de una persona.*”

Respuesta Humana: Preferencia

Respuesta correcta: Identidad sexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones	
Sistema antiguo:	15. identidad sexual	MP - 16	16	45	83	77	205	
Sistema mejorado:	16. identidad sexual	P - 14	16	2	30	1	33	
Mejora:				0.00%	95.56%	63.86%	98.70%	83.90%

12) “*Orientación sexual de un individuo en relación a sus prácticas sexuales.*”

Respuesta Humana: Preferencia

Respuesta correcta: Identidad sexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones	
Sistema antiguo:	17. identidad sexual	P - 24	31	7	57	41	105	
Sistema mejorado:				5	51	1	57	
Mejora:				N. E.	28.57%	10.53%	97.56%	45.71%

13) “Actuar de forma honesta y sincera.”

Respuesta Humana: Buen pedo

Respuesta correcta: Identidad sexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	18. moral	MP - 22	22	42	61	65	168
Sistema mejorado:	19. moral	P - 12	13	1	19	12	32
		Mejora:	40.91%	97.62%	68.85%	81.54%	80.95%

14) “Enfermedad venérea.”

Respuesta Humana: Gonorrea

Respuesta correcta: Herpes

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	20. Ampullitis	MP - 6	6	56	132	0	188
	21. blenorragia	MP - 11	11				
	22. contagio de enfermedades	MP - 14	14				
	23. ets	MP - 22	22				
	24. gonorrea	MP - 24	24				
	25. herpes genital	MP - 26	26				
	26. sífilis	MP - 49	49				
	27. transmisión sexual	MP - 52	52				
	28. analgia	P - 11	67				
	29. chancro	P - 25	81				
	30. cistitis	P - 27	83				
	31. sida	P - 113	169				
	32. tricomoniasis	P - 122	178				
	33. uretritis	P - 124	178				
	34. vih	P - 126	180				

Sistema mejorado:	35.	ampullitis	MP - 2	2	39	146	0	185
	36.	blenorragia	MP - 8	8				
	37.	contagio de enfermedades	MP - 9	9				
	38.	ets	MP - 15	15				
	39.	gonorrea	MP - 17	17				
	40.	herpes genital	MP - 18	18				
	41.	sífilis	MP - 32	32				
	42.	transmisión sexual	MP - 35	35				
	43.	analgia	P - 12	66				
	44.	chancro	P - 27	68				
	45.	cistitis	P - 29	164				
	46.	sida	P - 125	174				
	47.	tricomoniasis	P - 135	176				
	48.	uretritis	P - 137	179				
	49.	vih	P - 140					
Mejora:				11.36%	30.36%	-10.61%	N. E.	1.60%

15) “Enfermedad transmitida por el contacto sexual.”

Respuesta Humana: ETS

Respuesta correcta: Herpes

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	50. blenorragia 51. contagio de enfermedades 52. herpes genital 53. sífilis 54. transmisión sexual 55. ampullitis 56. ets 57. gonorrea 58. sida 59. uretritis 60. cistitis	MP - 8 MP - 9 MP - 12 MP - 24 MP - 27 P - 4 P - 16 P - 21 P - 46 P - 50 PP - 10	8 9 12 24 27 32 44 49 74 78 92	28	54	77	159
Sistema mejorado:	61. blenorragia 62. contagio de enfermedades 63. herpes genital 64. transmisión sexual 65. ampullitis 66. ets 67. gonorrea 68. sífilis	MP - 1 MP - 2 MP - 3 MP - 6 PP - 1 PP - 3 PP - 4 PP - 6	1 2 3 6 11 13 14 16	6	4	6	16
Mejora:			47.61%	78.57%	92.59%	92.21%	89.94%

16) “Enfermedad de transmisión sexual que produce daños fisiológicos en los genitales.”

Respuesta Humana: Sífilis

Respuesta correcta: Herpes

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	69. sífilis 70. herpes genital 71. ampullitis	MP - 3 P - 10 PP - 4	3 14 25	4	17	41	62
Sistema mejorado:	72. herpes genital 73. sífilis 74. ampullitis	MP - 6 MP - 7 PP - 1	6 7 9	7	1	12	20
Mejora:				47.61%	-75.00%	94.12%	70.73%
							67.74%

17) “Enfermedad infecciosa de transmisión sexual.”

Respuesta Humana: Un chingo

Respuesta correcta: Herpes

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	75. blenorragia 76. gonorrea 77. herpes genital 78. sífilis 79. ampullitis	MP - 1 MP - 3 MP - 4 MP - 6 PP - 9	1 3 4 6 37	6	22	86	114
Sistema mejorado:	80. blenorragia 81. gonorrea 82. herpes genital 83. sífilis 84. ampullitis	MP - 1 MP - 5 MP - 6 MP - 9 PP - 2	1 5 6 9 16	10	4	30	44
Mejora:				27.45%	-66.67%	81.82%	0,92
							61.40%

18) “Enfermedad de transmisión sexual.”

Respuesta Humana: SIDA

Respuesta correcta: Herpes

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones	
Sistema antiguo:	85. blenorragia 86. ets 87. gonorrea 88. herpes genital 89. sífilis 90. ampullitis 91. chancro 92. cistitis	MP - 4 MP - 7 MP - 9 MP - 10 MP - 18 P - 10 PP - 17 PP - 20	4 7 9 10 18 31 128 131	21	90	105	216	
Sistema mejorado:	93. blenorragia 94. ets 95. gonorrea 96. herpes genital 97. sífilis	MP - 1 MP - 3 MP - 5 MP - 6 MP - 9	1 3 5 6 9	10	1	4	15	
Mejora:				50%	52.38%	98.89%	0,92	93.06%

19) “Preservativo sexual del tipo de barrera.”

Respuesta Humana: Condón

Respuesta correcta: Condón

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones	
Sistema antiguo:	98. preservativo 99. condón 100. diafragma 101. capuchón cervical	MP - 3 P - 4 P - 6 PP - 18	3 7 9 56	3	35	101	139	
Sistema mejorado:	102. preservativo	MP - 1	1	1	1	1	3	
Mejora:				66.67%	66.67%	97.14%	99.01%	97.84%

20) “Método anticonceptivo fácil de usar y de adquirir.”

Respuesta Humana: Condón

Respuesta correcta: Condón

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	103. preservativo 104. condón	MP - 6 PP - 11	6 74	7	56	68	131
Sistema mejorado:				1	2	7	10
Mejora:			N.E.	85.71%	96.43%	89.71%	92.37%

21) “Método anticonceptivo de látex que cubre el cuerpo del pene para evitar una fecundación durante la relación sexual.”

Respuesta Humana: Condón

Respuesta correcta: Condón

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	105. preservativo	MP - 1	1	1	1	5	7
Sistema mejorado:	106. preservativo	MP - 2	2	2	2	1	5
Mejora:			-100.00%	-100.00%	-100.00%	80.00%	28.57%

22) “Barrera, generalmente de látex, que cubre el pene y evita el paso del semen hacia el útero en una relación sexual.”

Respuesta Humana: Condón

Respuesta correcta: Condón

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	107. preservativo 108. condón	MP - 1 PP - 2	1 6	1	3	9	13
Sistema mejorado:	109. preservativo	P - 1	2	1	1	2	4
Mejora:			-100.00%	0.00%	66.67%	77.78%	69.23%

23) “Método anticonceptivo.”

Respuesta Humana: Condón

Respuesta correcta: Condón

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	110. anticoncepción de emergencia	MP - 1	1	12	163	0	175
	111. anticonceptivo	MP - 2	2				
	112. capuchón cervical	MP - 3	3				
	113. diafragma	MP - 4	4				
	114. dispositivo intrauterino	MP - 5	5				
	115. espermicida	MP - 7	7				
	116. esponja espermicida	MP - 8	8				
	117. preservativo	MP - 10	10				
	118. píldora anticonceptiva	MP - 12	12				
	119. abstinencia sexual	P - 4	16				
	120. contracepción hormonal	P - 31	43				
	121. control natural de la natalidad	P - 33	45				
	122. crema espermicida	P - 34	46				
	123. cápsula cervical	P - 39	51				
	124. espermicidas vaginales	P - 51	63				
	125. método del ritmo	P - 91	103				
	126. método de la barrera	P - 92	104				
127. vasectomía	P - 154	166					

Sistema mejorado:	128. anovulatorio	MP - 6	6	60	0	0	60
	129. anticonceptivo	MP - 7	7				
	130. capuchón cervical	MP - 8	8				
	131. crema espermicida	MP - 14	14				
	132. diafragma	MP - 18	18				
	133. dispositivo intrauterino	MP - 19	19				
	134. espermicida	MP - 23	23				
	135. esponja espermicida	MP - 24	24				
	136. método del ritmo	MP - 40	40				
	137. método de la barrera	MP - 41	41				
	138. preservativo						
	139. píldora anticonceptiva	MP - 45 MP - 47	45 47				
	140. vasectomía	MP - 57	57				
	Mejora:						

24) “Cuando a alguien le gustan personas de ambos sexos es”

Respuesta Humana: Bisexual

Respuesta correcta: Bisexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	141. bisexual	MP - 1	1	1	30	68	99
Sistema mejorado:	1. bisexual	MP - 1	1	1	4	2	7
Mejora:			0.00%	0.00%	86.67%	97.06%	92.93%

25) “Atracción hacia ambos sexos.”

Respuesta Humana: Bisexualidad

Respuesta correcta: Bisexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	142. bisexual	MP - 2	2	8	48	132	188
Sistema mejorado:	1. bisexual	MP - 1	1	1	30	129	160
Mejora:				50.00%	87.50%	37.50%	2.27%
							14.89%

26) “Cuando a una persona le atraen personas de ambos géneros.”

Respuesta Humana: Bisexualidad

Respuesta correcta: Bisexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	143. bisexual	MP - 1	1	1	73	78	152
Sistema mejorado:	1. bisexual	MP - 1	1	1	5	21	27
Mejora:				0.00%	0.00%	93.15%	73.08%
							82.24%

27) “Atracción física hacia sexo de tipo masculino como femenino.”

Respuesta Humana: Heterosexual

Respuesta correcta: Bisexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	144. bisexual	P - 3	5	2	15	40	57
Sistema mejorado:	1.			2	5	10	17
Mejora:				N. E.	0.00%	66.67%	75.00%
							70.18%

28) “Atracción física hacia sexo de tipo masculino como femenino.”

Respuesta Humana: Bisexualidad

Respuesta correcta: Bisexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	145. bisexual	MP - 1	1	10	20	61	91
Sistema mejorado:	1.			1	1	1	3
		Mejora:	N. E.	90.00%	95.00%	98.36%	96.70%

29) “El gusto y afinidad amorosa y/o sexual por ambos géneros.”

Respuesta Humana: Bisexualidad

Respuesta correcta: Bisexual

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	146. bisexual	MP - 2	2	9	39	49	97
Sistema mejorado:	1.			1	5	6	12
		Mejora:	N. E.	88.89%	87.18%	87.76%	87.63%

30) “Familia / grupo, hay masculino y femenino.”

Respuesta Humana: Heterosexualidad

Respuesta correcta: Género

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	147. género	MP - 18	18	50	56	70	176
Sistema mejorado:	1.			2	2	2	6
		Mejora:	N. E.	96.00%	96.43%	97.14%	96.59%

31) “Clasificación con la que se distinguen los participantes en la reproducción sexual. Hay masculino y femenino.”

Respuesta Humana: sexo

Respuesta correcta: Género

	Mejores respuestas:	Posiciones relativas	Posiciones Absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	148. sexo	P - 11	12	1	12	36	49
Sistema mejorado:	1. sexo	P - 2	3	1	2	7	10
Mejora:				75.00%	0.00%	83.33%	80.56%
79.59%							

32) “Rasgo que permite distinguir el sexo al que pertenece una persona, animal o cosa.”

Respuesta Humana: género

Respuesta correcta: Género

	Mejores respuestas:	Posiciones relativas	Posiciones Absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	149. género	PP - 24	76	17	35	50	102
Sistema mejorado:				2	1	1	4
Mejora:				N. E.	88.24%	97.14%	98.00%
96.08%							

33) “Enfermedad donde las defensas del cuerpo son muy bajas. Se contagia por medio de relaciones sexuales.”

Respuesta Humana: SIDA

Respuesta correcta: SIDA

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:				2	14	21	37
Sistema mejorado:				1	2	2	5
Mejora:				N. E.	50.00%	85.71%	90.48%
86.49%							

34) “Inmuno deficiencia que afecta al hombre.”

Respuesta Humana: SIDA

Respuesta correcta: SIDA

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	150. SIDA	MP - 1	1	1	15	143	159
Sistema mejorado:	1. SIDA	P - 3	4	1	3	14	18
Mejora:			-300.00%	0.00%	80.00%	90.21%	88.68%

35) “Síntomas de la enfermedad causada por el VIH.”

Respuesta Humana: SIDA

Respuesta correcta: SIDA

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	151. SIDA	P - 71	85	14	86	120	220
Sistema mejorado:	1. SIDA	P - 5	8	3	6	16	25
Mejora:			90.59%	78.57%	93.02%	86.67%	88.64%

36) “Enfermedad de tipo sexual.”

Respuesta Humana: gonorrea

Respuesta correcta: SIDA

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	152. gonorrea 153. herpes genital 154. sífilis 155. ets 156. chancro 157. SIDA	MP - 17 MP - 19 MP - 39 P - 47 PP - 17 PP - 83	17 19 39 92 156 222	45	94	108	247
Sistema mejorado:	1. gonorrea 2. herpes genital	MP - 4 MP - 5	4 5	12	3	27	42
Mejora:			75.00%	73.33%	96.81%	75.00%	83.00%

37) “Siglas para el síndrome de inmunodeficiencia adquirida.”

Respuesta Humana: SIDA

Respuesta correcta: SIDA

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	158. SIDA	P - 12	16	4	13	160	177
Sistema mejorado:	1. SIDA	MP - 1	1	2	29	0	31
Mejora:			93.75%	50.00%	-123.08%	N. E.	82.49%

38) “Enfermedad de transmisión sexual incurable y altamente contagiosa.”

Respuesta Humana: SIDA

Respuesta correcta: SIDA

	Mejores respuestas:	Posiciones relativas	Posiciones absolutas	No. de opciones en “Muy probable”	No. de opciones en “Probable”	No. de opciones en “Poco probable”	No. total de opciones
Sistema antiguo:	159. SIDA	P - 21	24	3	25	89	117
Sistema mejorado:				10	6	1	17
		Mejora:	N. E.	-233.33%	76.00%	98.88%	85.47%