UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

COLEGIO DE LETRAS HISPÁNICAS

RECUPERACIÓN DE INFORMACIÓN EN DIRECTORIOS COMERCIALES WEB A PARTIR DE HERRAMIENTAS LINGÜÍSTICAS

INFORME ACADÉMICO POR ACTIVIDAD PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE LICENCIADA EN

LENGUA Y LITERATURAS HISPÁNICAS PRESENTA

WENDY FITZ ÁVILA

ASESORA

DRA. MARÍA DE LOS ÁNGELES ADRIANA ÁVILA FIGUEROA

MÉXICO D.F. 2011

Dedicatoria

A mis padres, mis hermanas, siempre incondicionales

A mis amigos,

A RedCúbica,

A mis sinodales

A los que se han quedado a mi lado

Y a los que se han ido...

Gracias

Cuántas palabras, cuántas nomenclaturas para un

mismo desconcierto

Rayuela, Julio Cortazar

Índice

Índice de figuras y tablas

Índice de figuras y tablas

Introducción

Las *nuevas tecnologías* propiciaron el inicio de la *sociedad de la información*. Éstas agrupan a las telecomunicaciones, la informática, la microelectrónica, el entorno multimedia y los medios electrónicos que sirven para almacenar, recuperar y transmitir la información de manera rápida y en gran volumen. La finalidad de estas tecnologías es la creación de canales de comunicación que respondan a las necesidades de la sociedad actual. De entre todos los anteriores, sin duda Internet y la web son los avances tecnológicos que han marcado el paradigma de los últimos años. Estos dos términos suelen llegar a confundirse: mientras que Internet es una combinación de *hardware* (conexiones físicas vía telefónica o satelital que posibilitan la interconexión entre todas las computadoras del mundo) con *software* (protocolos de comunicación y lenguajes que interpretan los comandos enviados por esa conexión física), la web es un sistema de comunicación y de publicación diseñado para distribuir información a través de esa misma red; su función principal es la de guardar información y presentarla al usuario cuando éste la requiera.

A principios de la década de los 90 estas tecnologías se masificaron, lo que significó el inició de una revolución en el campo de las comunicaciones, la educación, el comercio y los negocios. En la actualidad, Internet y la web se han vuelto una necesidad en la vida de cualquier persona en el mundo, más allá de si se está de acuerdo o no con este tipo de tecnologías y las implicaciones sociales que conllevan. Sin embargo, el exceso de información en la web ha reducido su potencial, lo cual ha supuesto un reto en cuanto a cómo organizar el conocimiento humano y optimizar un recurso de esta naturaleza. Más allá de esta cuestión, surge un debate sobre la libertad de expresión que ejercen los usuarios y sobre la propiedad intelectual de los contenidos existentes. En este punto, se debe recordar que el objetivo primordial de Internet y la web fue el intercambio de información libre de lucro entre investigadores y estudiantes; al respecto, se deberán redefinir los alcances y límites permitidos en un medio que desde sus orígenes se ha caracterizado por la libertad para consultar fuentes de referencia académicos o contenidos multimedia.

Internet y la web han contribuido a que la gente se relacione y comparta información de diversa naturaleza alrededor del mundo. En consecuencia, se puede afirmar que las fronteras geográficas están desapareciendo, no así las que implica la diversidad lingüística. Debido a la cantidad de información que almacena, muchos han pensado convertir a la web en la *Alejandría* de nuestros tiempos. Esta nueva biblioteca sólo alcanzará tan anhelado fin si se logra hacer confiable la información almacenada a través de una organización del conocimiento, construido a partir de ontologías y redes léxico-semánticas que permitan la estructuración de los documentos existentes de manera coherente. En este sentido, la web ha funcionado

hasta el momento más como un repositorio de información que como un sistema *inteligente*, esto es, que sea capaz de entender y resolver problemas por sí mismo. Se entenderá el papel fundamental que la lingüística juega para lograr esa meta a través de la inclusión de herramientas lingüísticas como ontologías, redes léxico-semánticas o lematizadores a los algoritmos de recuperación de información.

Para aquellos que se han formado bajo la cultura del libro y la biblioteca esta idea parece una aberración. No obstante, esta nueva realidad se alza en un territorio virtual en el que los estantes han sido sustituidos por *URL's* [1] (*uniform resource locator*, nombres que identifican a las páginas web*)* que identifican cada página en la web, y en el que los libros son reemplazados por páginas interminables gracias a los *links*[2] (enlaces) que las unen.

La evolución natural de los procesos de aprendizaje demuestra la oportunidad que representan estas tecnologías en la transmisión de conocimiento. Actualmente, algunas universidades en México han implementado programas de estudios a distancia, que permiten a los estudiantes acceder a sus clases desde cualquier lugar que disponga de una conexión a Internet; algunas de estas instituciones son la Universidad Nacional Autónoma de México, el Instituto Politécnico Nacional, el Instituto Tecnológico y de Estudios Superiores de Monterrey, entre otras. Por otro lado, las empresas privadas enfocadas a la gestión y recuperación de información web han incluido en sus equipos de trabajo a profesionales especializados en cuestiones lingüísticas. Tal fue el caso de la empresa RedCúbica S.A de C.V.

Este informe académico describirá el trabajo desarrollado en RedCúbica S.A. de C.V., durante los dos años del proyecto. En un principio, el objetivo de la empresa fue la creación de un directorio comercial en la web, semejante a *Sección amarilla*, pero que integrara factores semánticos para una mejor recuperación de información (RI). Se decidió replantear las metas a lograr a partir de un estudio de mercado en el que se analizaron las oportunidades de negocio que presentaba la comercialización de software lingüístico, tanto para el sector gubernamental como para la industria privada.

Los objetivos de este reporte son:

- Reseñar el origen y desarrollo de Internet, web y buscadores
- Analizar el papel de la lingüística en el crecimiento de las telecomunicaciones
- Describir las tareas realizadas dentro de la empresa RedCúbica S.A de C.V

[1] Una *URL*(Uniform resource locator) es la cadena de caracteres con la cual se asigna una dirección única a cada uno de los recursos de información disponibles en Internet; comúnmente es conocida como la dirección que tiene cada página web y se encuentra en la parte superior de la ventana del navegador utilizado para explorar la red. (Eubca, diccionario de organización del conocimiento. http://www.eubca.edu.uy/kod/espaniol/index.php)

[2] Se llama *Link* al elemento de una página web que la vincula con otra; mediante un *click* en el mencionado elemento se puede "explorar" la web de página en página. (*ídem*)

En el capítulo 1 se resumirán las características y el funcionamiento de los directorios comerciales impresos como antecedentes de los directorios *on line*; se revisará su importancia socioeconómica y se comentarán las deficiencias existentes. Posteriormente, y para contextualizar el desarrollo del tema, se hará una semblanza general acerca de Internet, web y buscadores. Por último, se hará una exposición de la evolución tecnológica que han tenido los directorios comerciales y los retos a los que se enfrentan en la web, además de su funcionamiento y los errores más comunes.

El capítulo 2 se centrará en aclarar el papel de la lingüística en el desarrollo de las nuevas tecnologías; se revisarán conceptos como *sociedad de la información*, *ingeniería lingüística*, *procesamiento del lenguaje natural* y *representación del conocimiento*. También se expondrá de manera general el funcionamiento de los recursos lingüísticos más usados para la recuperación de información.

Finalmente, en el capítulo 3 se describirán las metas logradas dentro de la empresa RedCúbica S.A de C.V. Asimismo, se describirá la metodología para el análisis y desarrollo de nuevas categorías, la organización de datos y el diseño de herramientas lingüísticas, como el lematizador, analizador morfológico y semántico, además de la integración de recursos léxicos a la propuesta integral del sistema de análisis lingüístico. Como parte del trabajo, se incluye la documentación realizada por el equipo lingüístico para establecer la metodología a utilizar en la creación de las categorías del directorio de la empresa. En el apéndice 1 se establecen las reglas sobre nominación de categorías y el apéndice 2 revisa los procedimientos para crear categorías a partir del análisis de otros directorios.

Este reporte académico tiene como finalidad la de señalar la falta de materias en los planes de estudio que vinculen la relación entre lengua y tecnología. En este sentido, la incursión laboral de un lingüista debería estar sustentada en conocimientos adquiridos previamente en las aulas de la facultad. Si bien es cierto que se ha dado mayor apoyo a la investigación de estos temas, la poca difusión de los mismos ocasiona que pocos alumnos participen en ellos. Para que México sea una potencia en el desarrollo de software lingüístico son necesarios más lingüistas interesados en el área.

Capítulo 1: Directorios comerciales en la web y recuperación de información

La empresa RedCúbica S.A de C.V tuvo como meta el desarrollo de un directorio comercial, tipo *Sección amarilla* que integrara factores semánticos y de lenguaje natural a los algoritmos de búsqueda creados para tal fin por los ingenieros y programadores que participaron en el proyecto. La inclusión de dichas

características contribuiría a subsanar la problemática que representaba (y sigue representando) el exceso de información en la web. En el año 2005, el desarrollo de software lingüístico era totalmente nuevo en México, no así en otras partes del mundo (Microsoft, Google, Teragram en Estados Unidos, Bitext en España, sólo por mencionar algunos proyectos similares); en este sentido, es destacable la visión emprendedora que tuvo la empresa RedCúbica S.A. de C.V. para su tiempo. Las actividades realizadas como lingüista consistieron en el análisis del funcionamiento de directorios comerciales web, así como de propuestas que ayudaran en la recuperación de la información. La finalidad fue ofrecer a los usuarios un menor tiempo en sus búsquedas gracias a la inclusión de herramientas de la lengua que proporcionaran mayor calidad en los resultados arrojados. En este capítulo se explicarán los antecedentes de los directorios comerciales web, así como el entorno tecnológico en el que se han implementado y los desafíos que enfrentan en cuanto a la recuperación de la información.

1.1 Antecedentes de los directorios comerciales web

Los nuevos medios de comunicación han impuesto a los tradicionales la necesidad del cambio y la adaptación. Hoy la televisión, la radio o los periódicos replantean sus estrategias de comercialización, pero también de difusión de sus contenidos. En el caso particular de los directorios comerciales web, los ajustes hechos han sido escasos y por lo tanto insuficientes para funcionar plenamente en un medio como éste. Para entender la problemática a la cual se enfrentan se revisarán cuáles eran los objetivos originales de estos catálogos en su versión impresa y cómo fueron transformándose hasta lo que es hoy una referencia comercial de primera mano.

1.1.1 Directorios comerciales impresos

Los directorios comerciales, mejor conocidos en el mundo como las *Páginas amarillas*, han tenido una larga tradición en el contexto económico mundial. Prácticamente se iniciaron con la aparición del teléfono. En un principio, estos directorios consistían solamente en un listado de números telefónicos de negocios de una determinada zona de la ciudad. Con el paso del tiempo se incluyeron anuncios de mayor formato y diseño, lo que transformó a estos catálogos en un canal esencial para la estrategia de ventas de los comerciantes; se puede afirmar que los directorios comerciales han sido una fuente de referencia que evidencia el desarrollo socioeconómico de las sociedades.

Víctor Cuchi, especialista en Historia económica de México, miembro de la Asociación Mexicana de Historia Económica e investigador de la UNAM, en su libro *Nacimiento de un mercado especial. El servicio*

telefónico en la ciudad de México, 1881-1911[3] reseña cómo inició el negocio de los directorios telefónicos en México. Refiere que en 1882, la Compañía Telefónica de México anunció la apertura de una central telefónica en la Ciudad de México. Este anuncio venía acompañado de una promesa: la tecnología les traería a los dueños de pequeños y grandes negocios ventajas tanto económicas como de logística pues tendrían mayor contacto con proveedores y distribuidores, además de obtener acceso a diversos servicios urbanos. La apertura de esta central telefónica fue compatible con el ideal del empresario cosmopolita de la época. El objetivo de la Compañía Telefónica de México era satisfacer la logística entre los negocios para agilizar las relaciones comerciales entre ellos. Incorporaba a la idiosincrasia del comerciante capitalino el ideal de trabajo dentro de una empresa. En sus inicios, la constitución del directorio telefónico coincidió con la apertura de grandes empresas que ya integraban a sus operaciones los servicios de telefonía como estrategia de comunicación interna. Para 1891, el directorio de la Ciudad de México contaba con 275 suscriptores; en 1902 esta cifra ascendió a 577 y para 1910 los suscriptores registrados fueron 2432[4]. Es pues, la necesidad y la estrategia de comercialización los que definieron el rumbo que tomarían estos directorios dentro de la vida de la sociedad mexicana.

Actualmente, la *Sección amarilla* impresa es distribuida por la compañía Teléfonos de México (TELMEX); como ya se ha mencionado, existen ediciones estatales y zonales, en ambas lo único que cambia son los anunciantes, el índice de clasificaciones es el mismo para todos los estados de la República Mexicana. Para efectos del presente trabajo sólo se tomó en cuenta el directorio de la zona Santa Fe-Polanco (zona Poniente).

1.1.1.1 Características

A continuación se presentarán las características principales de estos directorios:

- Extensión: los directorios son libros de más de 1000 páginas, cuya característica principal es el color amarillo de sus páginas. Asimismo, existen ediciones reducidas en su contenido que sólo muestran una zona determinada de la ciudad. En el caso específico de la Ciudad de México estos se dividen en zona sur, norte, oriente y poniente.

[3] Cuchi, Víctor. "Nacimiento de un mercado especial. El servicio telefónico en la Ciudad de México, 1881-1911" [en línea]. *Munich Personal RePEc Archive*, núm. 1027, 2007. http://mpra.ub.unimuenchen.de/1027.pdf
[4] Compañía Telefónica Mexicana. *Directorio telefónico de la Ciudad de México, año de 1891.* México: Condumex, 1991.

- Distribución: se distribuyen gratuitamente en todas las colonias de la República Mexicana; los distribuidores regalan tanto las versiones estatales, en la que se publican todos los negocios del estado, y las zonales, en las que se anuncian únicamente comerciantes locales.
- Índices y categorías: tienen un índice de clasificaciones que organiza la información de acuerdo con categorías (aproximadamente seiscientas). Estas categorías se pueden dividir entre las que anuncian a profesionistas (*dentistas*, *maestros*, *abogados*), la que ofrecen servicios (*mensajería*, *albañilería*, *plomería*, *contaduría*), ubican lugares (*escuelas*, *hospitales*, *restaurantes*) y venden productos (*teléfonos*, *leche*, *ropa*). La extensión de estos índices es larga, debido a la duplicidad de categorías que clasifican un mismo producto o servicio bajo distintos nombres, por ejemplo: *academias de canto* y *escuelas de enseñanza artística.*
- Tipo de anuncio: los anuncios pueden ser en formato sencillo o abarcar hasta una página entera, dependiendo de las expectativas mercadotécnicas de los anunciantes. Lo anterior representa una competencia desigual entre aquellos anunciantes que puedan pagar más y los que sólo quieren ser reconocidos dentro de este medio de gran tradición. Como siempre, los intereses comerciales privan más que la justa competencia del mercado. Las cotizaciones varían, dependiendo del formato del anuncio (un cuadro, un cuarto de página, media página o página completa)y de los colores usados (blanco y negro o a color). La página más cara es de 1 000 000 de pesos por una hoja completa a color. Empresas como Palacio de Hierro, Liverpool, Hertz y LaSalle tienen este tipo de publicidad en el directorio.

1.1.1.2 Funcionamiento y problemática

Estos catálogos de información tienen una organización semejante a la de un diccionario. Al igual que éstos, los directorios se organizan alfabéticamente a partir de la primera palabra que le da nombre a la categoría. En el contexto de los directorios comerciales, se entenderá por categoría al nombre genérico que agrupa un conjunto de artículos con características similares, relacionados entre sí. La inclusión o no de determinados artículos dependerá del grado en que puedan reconocerse, diferenciarse y en dado caso clasificarse como tales dentro de una categoría. El primer paso es la búsqueda de una categoría en específico dentro del índice de clasificaciones, para localizarla posteriormente en el directorio.

La búsqueda de las categorías es problemática debido a factores como la duplicidad de las mismas (*escuelas de enseñanza artística* y *escuelas de canto*) o al criterio arbitrario usado para clasificarse (excesiva libertad en la elección de palabras clave por parte de los anunciantes y falta de supervisión de los revisores de contenidos de *Sección amarilla* para verificar la información proporcionada, ver apéndice 2 para mayor ilustración del problema referido); en este sentido, los anunciantes tienen la libertad de elegir las categorías en las cuales desea ser publicado. Muchas veces, eligen categorías que no tienen relación con su giro comercial, lo cual en cierto modo es fomentado por criterio de *Sección amarilla* que propicia esta situación: "todos creemos que un anuncio de Hotel debe estar en la clasificación de *hoteles*, sin embargo, a veces es mejor anunciarse en *turismo* o *aerolíneas* porque generalmente quien busca una línea aérea, también necesita de un *hotel*"[5]. Como se apreciará en la *Sección amarilla* priva el interés comercial sobre el beneficio del usuario final, que busca rapidez y no información extra.

1.2 Entorno tecnológico: Internet, web y buscadores

Para una mejor comprensión del objetivo de este reporte académico, se ha considerado explicar el entorno dentro del cual se desarrollan los directorios comerciales. Este marco referencial ayudará en la interpretación del mismo en la medida en que pueda esclarecer el funcionamiento y los objetivos tecnológicos implícitos en su desarrollo.

1.2.1 Internet

[5] Adame, Goddard, Lourdes. "Sección amarilla: un impulsor estratégico de la telefonía comercial en México" [en línea]. *Instituto Mexicano de Teleservicios* núm. 22, 2008. http://www.imt.com.mx/rcforum/22/seccionamarilla.php

1.2.1.1 Definición

Resulta especialmente difícil dar una sola definición sobre Internet. Desde el punto de vista tecnológico se trata de una infraestructura de redes que conecta millones de computadoras en el mundo y que utiliza fibra óptica, satélites o telefonía convencional para tal fin. Formalmente es una red flexible, dinámica y adaptable, por lo que no se ajusta a ningún tipo de tecnología en específico ni a ninguna conexión o medio electrónico en concreto. Sin embargo, existen otros modos de apreciar el significado de Internet: desde el punto de vista social, esta red representa un cambio socioeconómico y cultural que facilita la interacción de las comunidades a partir de una realidad virtual. Lo anterior no supone la desaparición del contacto humano, pero si añade otras posibilidades de comunicación. Y es en este punto, el de la comunicación, en el que se aprecia su valor real: a diferencia de los medios tradicionales como la radio, la televisión o la prensa, Internet si permite una interacción en tiempo real entre los emisores de un mensaje y sus receptores. En Internet, un individuo es potencialmente emisor y receptor a la vez. En los siguientes apartados, se revisarán sus antecedentes, así como su funcionamiento, temas esenciales para una completa comprensión del tema tratado en este reporte académico.

1.2.1.2 Antecedentes

El avance tecnológico del siglo XX facilitó la cristalización de muchos de los proyectos iniciados en el pasado en torno a la comunicación y la conexión entre las sociedades. El más importante, no sólo por las implicaciones tecnológicas de su época, sino por su repercusión social fue Internet.

El deseo de hacer viable una red de conocimiento universal hizo posible el desarrollo de distintos proyectos para conseguir tan anhelado fin. Al respecto, para 1960, Licklider en su libro *Libraries of the future*[6] propuso *una red de muchos [ordenadores], conectados mediante líneas de comunicación de banda ancha las cuales proporcionarían las funciones hoy existentes de las bibliotecas junto con anticipados avances en el guardado y adquisición de información y otras funciones simbióticas.*

Mattelart [7] en su libro *Historia de la sociedad de la información* hace una amplia reseña sobre cómo fue que Internet nació en plena Guerra Fría. En 1957, el departamento de defensa del gobierno de Estados Unidos instituyó la *Advanced Research Projects Agency* (ARPA). Esta división se encargaría de asegurar el liderazgo de la nación en el campo de la ciencia y la tecnología. Para 1969, ARPA desarrolló la *Advanced Research Projects Agency Network* (ARPANET), red predecesora de Internet. Su invulnerabilidad residía en la

[6] Licklider, J.C.R.*Libraries of the future*. Cambridge, Mass: MIT Press, 1965.
[7] Mattelart, A. *Historia de la sociedad de la información 1ª ed.* Barcelona: Bolsillo Paidos, 2007.

eliminación de una autoridad central al dotar a cada computadora conectada a ARPANET la misma capacidad para mandar y recibir información. De esta manera, nadie que quisiera atacar el sistema de comunicación del gobierno de Estados Unidos podría saber por dónde empezar.

La primera de estas redes sin nodos centrales, basada en conmutación de paquetes, y el primer ordenador *host*[8] (servidor) se establecieron en la Universidad de California (UCLA). Ese mismo año cuatro servidores más estaban conectados a ARPANET; con ello se dio origen a lo que hoy es conocido como Internet.

Durante los siguientes años se desarrollaron la mayoría de los protocolos utilizados para la conexión entre computadoras. Estos protocolos son las reglas que las computadoras usan para comunicarse unas con otras a través de una red de comunicación. Estas normas se estandarizaron hasta lo que hoy se conoce como TCP/IP (*Transfer Control Protocol/Internetworking Protocol*); el objetivo de este protocolo es dividir los mensajes, dirigirlos al receptor y reconstruirlos una vez que han llegado.

Pronto diversas instituciones gubernamentales y privadas desarrollaron redes informáticas que se adaptaron al protocolo TCP/IP. Entre estas nuevas redes apareció la creada por la *National Science Foundation* (NSF). A esta red empezaron a asociarse otras; lo anterior y el incremento en su capacidad de transmisión de datos ocasionaron que ARPANET se declarara disuelta para 1989.

Sin embargo, el modelo original de Internet no tenía previsto alojar un gran número de redes. Era preciso un cambio tecnológico que definiera los tipos de redes que existirían (ya fueran nacionales, regionales y locales), así como de un sistema que asignara nombres de *dominio*[9] para identificarlas y resolver de forma jerárquica los nombres de los servidores en las que se alojarían cada una de las direcciones de Internet (por ejemplo .gob, .mx, .us).

María Jesús Lamarca comenta en *Hipertexto: el nuevo concepto de documento en la cultura de la imagen*[10], que en la actualidad Internet es una serie de servicios que tienen que ver con funciones de información, comunicación e interacción entre usuarios. Entre los servicios que ofrece Internet se

---

[8] El término *host* se refiere a la red de ordenadores conectados a la red; son los encargados de proveer o utilizar servicios de ella y se utilizan para tener acceso a la web. (Barité, Roqueta, Mario. "Diccionario de organización del conocimiento, clasificación, indización, terminología" [en línea]. Universidad de la República Oriental del Uruguay, 2011 http://www.eubca.edu.uy/kod/espaniol/index.php)

[9] Un dominio identifica a los servidores conectados a la red. Indica la "nacionalidad" del servidor que aloja la página web (si es mexicano su dominio será .mx, si es peruano será .pr, si es brasileño será .br). *ídem*

[10] Lamarca Lapuente, M.J. "Hipertexto: El nuevo concepto de documento en la cultura de la imagen". Tesis [en línea]. Universidad Complutense de Madrid, 2009. http://www.hipertexto.info/documentos/h_hipertex.htm

encuentran aquellos que permiten el intercambio de mensajes personales (correo electrónico, foros, blogs, etc.), los que ofrecen conversaciones en tiempo real (chats) y los dedicados al suministro y acceso a la información (*World Wide Web* y FTP [*File Transfer Protocol*]).

Estos servicios han evolucionado, perfeccionando la experiencia en su uso y manejo. Tecnológicamente han mejorado al permitir mayor velocidad de transmisión, mayor ancho de banda, mejoras en el software y el hardware (mayor almacenamiento de información, velocidad de procesadores, versatilidad en el uso). El usuario, a diferencia de los primeros años de Internet, ya no tiene que aprender comandos y algoritmos complejos sino simplemente manejar un ratón y posarlo sobre íconos e interfaces[11] gráficas, integrando las funciones del lenguaje natural (voz y texto).

1.2.1.3 Funcionamiento

Internet funciona a partir de protocolos de comunicación que permiten la interconexión e intercambio de información entre equipos con diversos sistemas operativos o de *hardware*. El protocolo estándar es el TCP/IP. Sus funciones son:

- Lograr una interconexión física mediante cableado telefónico convencional, fibra óptica, satélites, enlaces por microondas, entre otros. Este cableado *virtual* permite disponer de un canal de comunicación entre dos computadoras situadas en diferentes lugares de la red.
- Localizar una computadora dentro de Internet, con independencia de su situación física y de los enlaces de comunicaciones para poder alcanzarlo.
- Disponer de un lenguaje común de intercambio de información, entendido por todos ellos, y que sea independiente de su estructura interna o sistema operativo.
- Resolver problemas que se presenten durante el intercambio de datos
- Aclarar las posibles incompatibilidades en la comunicación entre ordenadores debidas a los diferentes sistemas de representación digital de la información que éstos utilizan.

1.2.2 Web

1.2.2.1 Definición

La web es el sistema de documentos interconectados por los enlaces de hipertexto disponibles en Internet. Como se ha hecho mención, estos dos términos no son sinónimos: aquel es sólo un servicio de los muchos

[11] Una interfaz es la cara visible de los programas informáticos. Abarca las pantallas y su diseño, el lenguaje de programación usado, entre otros aspectos de comunicación usados entre máquina/persona. *ídem*

que ofrece este último. Un usuario podrá visualizar sitios web que contengan texto, imágenes u otros elementos multimedia y navegará a través de ellos usando *links* (enlaces).

1.2.2.2 Antecedentes

Para hacer más eficiente el intercambio de comunicación entre los usuarios de Internet se pensó en la vieja idea del hipertexto propuesta por Vannevar Bush en los años 40. Un hipertexto es *una serie de elementos de información (desde fragmentos hasta documentos completos) que se relacionan entre sí*. Los autores del hipertexto crean los nodos[12] y los enlaces, y el lector del hipertexto puede *viajar* activando esos nodos.

Lamarca refiere que, en 1989, el *Centre Europeen de Recherche Nucleaire* (CERN) realizó varias pruebas para crear un sistema de comunicación entre los científicos nucleares de todo el mundo. Entre las pruebas que usaron para conectarse a Internet se estandarizó el protocolo de conexión TCP/IP. Tim Berners-Lee, ingeniero especializado en telecomunicaciones, propuso además un sistema de comunicación basado en la tecnología de hipertexto. Este proyecto permitía almacenar piezas de información multimedia (texto, video, sonido e imágenes), conectarlas y enviarlas a los usuarios finales. El objetivo de Berners-Lee fue el desarrollo de un método eficiente y rápido para intercambiar datos entre la comunidad científica. El resultado fue la combinación del protocolo de internet y el hipertexto. Lo anterior creó un nuevo acceso a la información, intuitivo e igualitario, lo que hace posible que cualquiera pueda acceder a Internet.

Esta tecnología se basa en el lenguaje HTML (*HyperText Markup Language*) y en el protocolo HTTP (*HyperText Transfer Protocol*). Este protocolo permite visualizar de forma rápida y sencilla todo tipo de información (video, audio, texto y software). La unión de estos enlaces se da mediante la URL. Lo anterior convierte a esta herramienta en un motor fundamental en el desarrollo de Internet. En la práctica, es el servicio más usado de Internet pues no sólo funciona como un depósito de información sino también como una forma de acceso, búsqueda y recuperación.

1.2.2.3 Funcionamiento

Para utilizar el servicio web es necesario un servidor y un cliente[13] (cuya relación se representa en la figura 1). El servidor es la empresa que proporciona las computadoras donde se alojan las páginas web codificadas bajo el estándar HTML; el cliente es la computadora del usuario. Si éste quiere acceder a esas páginas web tiene que utilizar un programa que interprete el código HTML. Estos programas son los

[12] Un nodo es el punto de unión entre varias redes; su importancia radica en que gestiona la rapidez de las conexiones de la computadora. *ídem*

navegadores, y su tarea es la de permitir al usuario visualizar los contenidos de la web al interpretar el lenguaje HTML; Los navegadores más usados son *Internet explorer*, *Netscape navigator*, *Crhome*, *Mozilla firexfox*, entre otros.

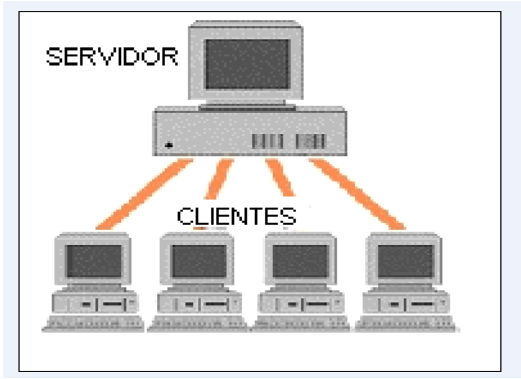


Figura 1. Representación de la relación entre servidor-cliente

Explica Mattelart que, en 1991, el programa creado por Berners-Lee se empezó a distribuir gratuitamente en el mundo académico; para 1994 había más de 50 millones de usuarios. En la actualidad, la *World Wide Web* es el servicio más utilizado de Internet y se considera la causa y vehículo principal de su éxito.

La gran contribución de Tim Berners-Lee fue unir un hipertexto a la información contenida en Internet. La Web reúne documentos y recursos de todo tipo, como imágenes, audio, videos, recursos audiovisuales.

1.2.3 Buscadores

1.2.3.1 Definición

Un buscador es una página web que explora los archivos almacenados en servidores web, a partir de palabras clave o árboles jerárquicos organizados por temas. El resultado es un listado de direcciones web en los que se mencionan contenidos relacionados con las palabras clave ingresadas en la caja de búsqueda.

1.2.3.2 Antecedentes

Lamarca reseña que, para 1990, el aumento de usuarios de Internet provocó que la NSF se hiciera cargo de la administración de la red. Internet comenzó a saturarse debido a que su objetivo no había sido ideado para las dimensiones de información que había alcanzado. Ante este panorama se desarrollaron los buscadores.

En abril de 1994 David Filo y Jerry Yang, universitarios, crearon una página web en la que se ofreciera un directorio de páginas interesantes clasificadas por temas, enfocadas principalmente a las necesidades de información que tuvieran estudiantes como ellos. Ese fue el inicio de *Yahoo!*, uno de los buscadores más populares en el mundo.

1.2.3.3 Funcionamiento

Los buscadores son las herramientas para localizar información en Internet. Su función es gestionar las bases de datos de las *URL's* obtenidas a través de la exploración e indización de los contenidos multimedia de la web. Existen distintos tipos:

- Los que gestionan catálogos o directorios por categorías, se conocen como índices temáticos.
- Los que analizan páginas web a partir de los temas tratados en ellas, se llaman motores de búsqueda.
- Los que combinan las potencialidades de varios motores de búsqueda, se llaman metabuscadores.

Los pasos que sigue un buscador son:

- Recogida y análisis de datos (indización y clasificación de categorías)
- Búsqueda propiamente dicha
- Recuperación de información

Mientras que los directorios por categorías lo construyen equipos de personas que exploran la red para encontrar páginas relacionadas a los temas de la clasificación, en los motores de búsqueda el trabajo de recogida y análisis de datos lo hace de manera automática un *spider* (araña) que rastrea la red y guarda la dirección *URL* para darla de alta en una base de datos, misma que servirá posteriormente para devolver resultados a los usuarios que consulten el buscador. Este robot analiza la página principal y todas las que se enlazan a ella, y lee etiquetas META (metadatos que contienen la descripción de la aplicación, así como su relación con otros documentos) para extraer los datos contenidos en ellas a través del lenguaje HTML. Con esa información, el robot indiza palabras clave como título, idioma, autor, propietario, localización, temas, etc.

Existen sistemas de búsqueda que mezclan estas dos funciones y ofrecen tanto búsquedas por medio de índices temáticos, como búsquedas libres por palabras clave. Un sistema ideal sería aquel que permitiría encontrar información por medio de esas dos posibilidades. Actualmente, un buscador contempla las búsquedas a través de palabras clave, pero también por medio de palabras truncadas, además de incluir operadores booleanos como *and*, *or* y *not* que indican o restringen la relación entre dos o más términos.

También incluyen la posibilidad para elegir el idioma. A últimas fechas, muchos motores de búsqueda han incluido factores lingüísticos con el fin de aclarar ambigüedades, sinonimias y polisemias.

En la figura 2 se muestra una página de resultados con páginas web referentes al tema *Cuisine.* Para que un motor de búsqueda pueda mostrar esos resultados primero se debieron definir las estrategias a utilizar para que una página web pudiera ser indizada por el motor de *Yahoo!*:

- Definición de las palabras clave con las que se pretenda identificar a la página web, éstas se incluirán en el título de la página y en el contenido de la misma (por ejemplo, las palabras clave para el sitio web de *la Comer* serían *miércoles de plaza, alimentos, promociones, ofertas, ahorro*, por citar algunas).
- La *URL* deberá estar entre esas palabras clave (siguiendo el ejemplo anterior la palabra clave sería *www.comercialmexicana.com.mx).*
- Incorporación de metadatos e información relevante en las primeras líneas de la página web, entre ellas las palabras clave (esta parte se refiere a la incorporación de esas palabras clave dentro del código de programación).
- Creación de *links* que referencien la página web (esto quiere decir que deberá enlazarse a otras páginas a fin de que el motor la encuentre, la índice y la presente en la página de resultados).

Figura 2. Representación del funcionamiento de una búsqueda por medio de palabras clave

Un buscador está compuesto por:

- Robot: son programas que recorren la Web y recuperan documentos de forma automática. De esta manera constituyen una base de datos con un listado de *URL's* que visitan periódicamente.

- Indexador: se trata de un programa que recibe las páginas recuperadas por el robot. Extrae la información y la transforma en una base de datos. Las técnicas para extraer información son:
    - ✓ Listas de *stop*: listas de palabras habituales que no aportan significado y que no deben aparecer en el vocabulario (preposiciones, conjunciones, artículos).
    - ✓ Extracción de raíces: consigue un lema único para palabras con formas parecidas.
- Motor de búsqueda: este programa se encarga de analizar una consulta de usuario y buscar en el índice los documentos relacionados. El motor de búsqueda será capaz de ordenar los resultados de manera que aparezcan antes las páginas más relevantes atendiendo a varios indicadores, como:
    - ✓ Localización: la página de resultados muestra los documentos en los cuales aparecen las palabras utilizadas en la consulta. Las páginas web más relevantes siempre aparecerán al inicio de la página de resultados.
    - ✓ Frecuencia de aparición: a mayor número de apariciones de los términos de la consulta en una página, más relevante será ésta para el resultado.
    - ✓ Popularidad: se mide en el número de enlaces que apuntan a esa página, esto es, las veces que otras páginas web hacen mención y enlazan a la página en cuestión. Una página a la que se hacen muchas referencias suele ser mejor que otra a la que se hacen menos.
- Interfaz: la más utilizada en la web son los formularios:
    - ✓ Formularios: generalmente están basados en una caja de texto (en donde el usuario introduce la palabra clave o frase buscada) y un botón de envío. Para búsquedas más avanzadas existen formularios más avanzados que permiten introducir varias palabras, añadir operadores booleanos, buscar en un idioma concreto, buscar por proximidad, entre otros.
    - ✓ Páginas web de resultados: es un listado en el que se muestran una lista de páginas web relacionadas a la búsqueda. Cada uno de éstas contiene una descripción, el contexto en el que se ha encontrado y el enlace para visualizarla completamente.

Los robots son programas que rastrean la estructura hipertextual de la web, recogen información sobre páginas, indizan información, la clasifican y conforman una base de datos que servirá para que los motores de búsqueda encuentren la información solicitada por los usuarios.

1.2.3.3.1 Recuperación de información

1.2.3.3.1.1 Definición

Se puede afirmar que un buscador no es nada en el mundo de la web si antes no ha definido su estrategia de recuperación de información (RI). La RI es parte fundamental de un buscador en cuanto a que define cómo se representará, organizará y difundirá la información obtenida a través de la indización de *URL's* obtenidas por el robot. Es una especie de armazón que sostiene la gran casa que representan los documentos en la web. Korfhage definió a la RI como *la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta*[14]. La RI es un área que abarca aspectos computacionales como el almacenamiento y la organización de datos, y los relacionados con la representación y recuperación de datos, que conciernen propiamente a los usuarios y el lenguaje natural que utilizan para hacer sus consultas.

1.2.3.3.1.2 Funcionamiento

La RI se puede estudiar desde dos puntos de vista: el computacional y el humano. Aquel se refiere al cómo se estructuran los datos y a los algoritmos utilizados para mejorar las respuestas, mientras que este último, el factor humano, estudia el comportamiento y las necesidades de los usuarios.

Para cumplir sus objetivos, la RI debe realizar tareas básicas, formuladas en cuestiones computacionales:

- Representación lógica de los documentos (cómo están organizadas las bases de datos).
- Representación de la necesidad de información del usuario en forma de consulta (cómo es que los usuarios del sistema harán una consulta.
- Evaluación de los documentos respecto a una consulta para establecer la relevancia de cada uno; a partir de esto se determinarán los documentos más relevantes que conformarán la respuesta *ideal* al usuario.

Como se observará, se parte de un conjunto de documentos de texto, compuestos por palabras que forman estructuras gramaticales (por ejemplo, oraciones y párrafos). Tales documentos están escritos en lenguaje natural y expresan ideas de un autor sobre determinado tema. El conjunto estos documentos se denominan *corpus*, *colección* o *base de datos textual* o *documental*. Para realizar operaciones sobre un corpus, es necesario obtener primero una representación lógica de todos sus documentos, la cual puede

[14] Tolosa, Gabriel. "Introducción a la recuperación de la información, conceptos, modelos y algoritmos básicos". Tesis [en línea]. Universidad Nacional de Luján, 2007. http://www.tyr.unlu.edu.ar/tallerIR/2008/docs/Introduccion-RI-v9f.pdf

consistir en un conjunto de términos, frases u otras unidades (sintácticas o semánticas) que los caractericen. Por ejemplo, la representación de los documentos mediante un conjunto de sus términos se la conoce como *bolsa de palabras* (*bag of words*).

A partir de la representación lógica de los documentos, existe un proceso de indexación que llevará a cabo la construcción de estructuras de datos (denominadas índices) que la almacene. Estas estructuras darán luego soporte para búsquedas eficientes mediante algoritmos. Éstos aceptarán como entrada una expresión de consulta de un usuario y verificará en el índice almacenado cuáles documentos pueden satisfacerlo. Luego, otro algoritmo determinará la relevancia de cada documento y retornará una lista con la respuesta. Se establece que el primer resultado de dicha lista corresponde al documento más relevante respecto de la consulta y así sucesivamente en orden decreciente. La interfaz de usuario permitirá que éste especifique la consulta mediante una expresión escrita en un lenguaje natural y – además – servirá para mostrar las respuestas retornadas por el sistema en ese mismo lenguaje. Si bien hasta aquí se planteó la tarea básica de la RI y de su arquitectura general, el área es muy amplia y abarca diferentes tópicos. En general, un sistema de RI no entrega una respuesta directa a una consulta, sino que permite localizar referencias a documentos que pueden contener información útil.

1.3 Directorios comerciales web

1.3.1 Definición

La evolución tecnológica supuso un reto para los directorios comerciales. Aunque muchos especialistas auguraron su desaparición, estas fuentes de referencia comercial han sabido sobrevivir en la era digital. Sin embargo, el ajuste de los contenidos al formato digital no ha sido del todo exitoso debido a que no se trata de una simple transcripción, sino de la adaptación de los contenidos a la dinámica de los usuarios para encontrar información en la web. Los usuarios que consultan el directorio *on line* buscan rapidez y eficiencia. Por lo tanto, las categorías deben ser sencillas y contemplar la inclusión de palabras clave y listas de sinónimos[15]; también deben tener en cuenta los problemas naturales de la lengua, como la polisemia o la homonimia. Los directorios comerciales son muy diferentes a sus pares impresos, no sólo por los contenidos que manejan sino por la forma en la que funcionan.

Figura 3. Imagen de la primera página del portal de *Sección amarilla*

[15] A lo largo del texto, deberá entenderse como *palabra clave* aquella expresión significativa compuesta por una o más palabras. Éstas pueden aparecer indistintamente dentro del texto o el título del mismo; asimismo, se entenderá por sinónimo a las palabras que tienen una misma o muy parecida significación que otro (DRAE, *sinónimo*)

1.3.2 Características

A continuación se describirán las características y los problemas que enfrentan estos directorios comerciales web:

- Han incorporado motores de búsqueda a sus sistemas de bases de datos privadas; de un tiempo a la fecha se ha contemplado la inclusión de herramientas lingüísticas, como correctores ortográficos, aunque de forma básica.
- Tienen las mismas clasificaciones que el libro impreso.
- Ofrecen la posibilidad de búsqueda por área geográfica.
- A partir de su incursión en la web, han diversificado su oferta de servicios, por lo cual han incluido oportunidades comerciales en nichos tecnológicos de moda (*mensajes SMS*, *GPS* y aplicaciones para *smartphones*[16]).

1.3.3 Problemática

El principal problema del directorio comercial web radica en ser una transcripción del impreso. A lo anterior, habría que añadir la falta de actualización del índice de clasificaciones y, quizás lo más importante, la incomprensión del proceso de búsqueda de información en la web. La inclusión de palabras clave que identifiquen plenamente a una categoría, junto a una buena selección de sinónimos, y la resolución de

[16] Los *mensajes SMS* (*Short Message Service*) son un servicio disponible en teléfonos celulares. El sistema *GPS* (*Global Positioning System*) permite determinar la posición de un objeto dentro de una zona específica.
Los *smartphones* son los llamados teléfonos "*inteligentes*" cuyas características los hacen similares a una computadora.

problemas relacionados a la ambigüedad propia de la lengua son claves en el éxito de estos directorios. A continuación se reseñan los principales problemas identificados y la forma en la que podrían resolverse.

1.3.3.1 Índice de clasificaciones

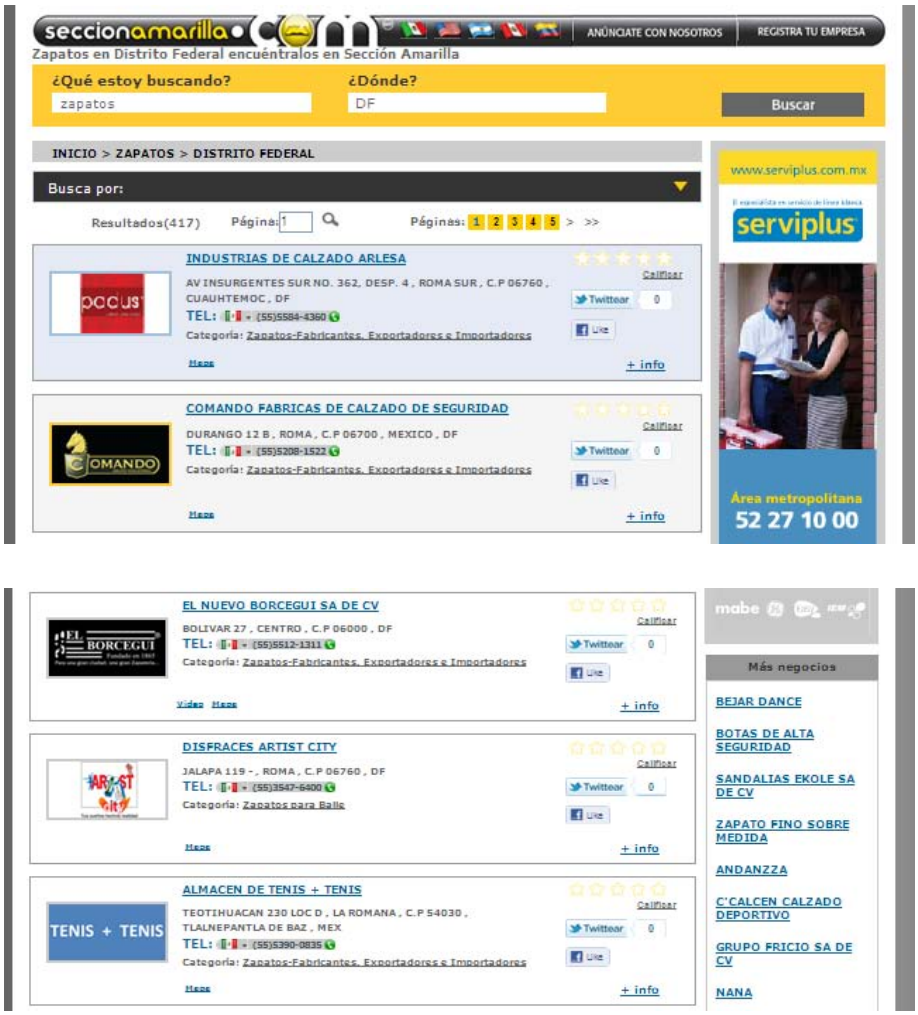
Uno de los mayores problemas de los directorios comerciales *on line* es la falta de actualización del índice de clasificaciones (véase apéndice 2 para mayor referencia). Esta *conformidad* con el modelo de negocio existente ha ocasionado que las categorías sean las mismas desde hace más de medio siglo. En la actualidad existen negocios que no encajan en las características comerciales de las categorías existentes en ese índice.

La duplicidad de categorías también representa un reto para estos directorios. Mientras que en el directorio impreso era entendible su inclusión debido principalmente al formato, en la versión web no hay justificación que explique la existencia de categorías duplicadas (éstas pueden incluirse como palabras clave y no como categorías en sí mismas). Otro problema es la libertad de la que gozan los anunciantes para elegir las categorías en las cuales quieran aparecer publicados, además de la libertad para elegir las palabras clave para que los usuarios los encuentren.

Esta falta de criterio y supervisión ocasiona una clasificación en categorías que no tienen relación con su giro comercial y en la elección de palabras clave y sinónimos indiscriminadamente, lo cual afecta los resultados de las búsquedas.

La supervisión de los revisores de contenidos es vital para que los anunciantes no abusen de la libertad que les da el formato web. Esto no debe interpretarse negativamente pues se debe tener en cuenta que se busca la rapidez y eficiencia que caracteriza (o debería de hacerlo) a la red. Como se apreciará en la figura 4, la palabra *zapatos* retorna resultados no relacionados directamente, como *zapatos, fabricantes, exportadores e importadores*. No se trata de adivinar lo que un usuario quiere cuando escribe tal o cual palabra, sino de entender el procedimiento cognitivo que realiza una persona cuando escribe una palabra: ¿qué es lo que realmente quiere encontrar? ¿sería conveniente, entonces, mostrarle resultados relacionados, aunque no todos fueran de su interés?

Figura 4. Representación de la búsqueda de *zapatos*



1.3.3.2 Palabras clave y sinónimos

Hasta el momento, la forma tradicional para encontrar información en la web es a través de palabras clave; esta situación se extiende a los directorios comerciales web. Un usuario escribe la palabra clave o el sinónimo, en todo caso, que él considere le pueda devolver más resultados relacionados a su criterio de búsqueda. Entre más acertada sea la elección de la palabra clave (o el sinónimo) más resultados encontrará. Sin embargo, lo anterior no se logra debido a que los anunciantes eligen palabras clave que no guardan relación directa a su giro comercial (por ejemplo, la palabra *verduras* para la categoría *carnicería*) o que pueden crear ambigüedad (por ejemplo, la palabra *gato* para las categorías *veterinario, mecánicos* o *servicio doméstico).*

1.3.3.3 Ambigüedad

Formalmente la ambigüedad se presenta cuando un término o estructura semántica se presta a una o más interpretaciones (DRAE, *ambigüedad*). En el contexto de los directorios comerciales web, la ambigüedad se

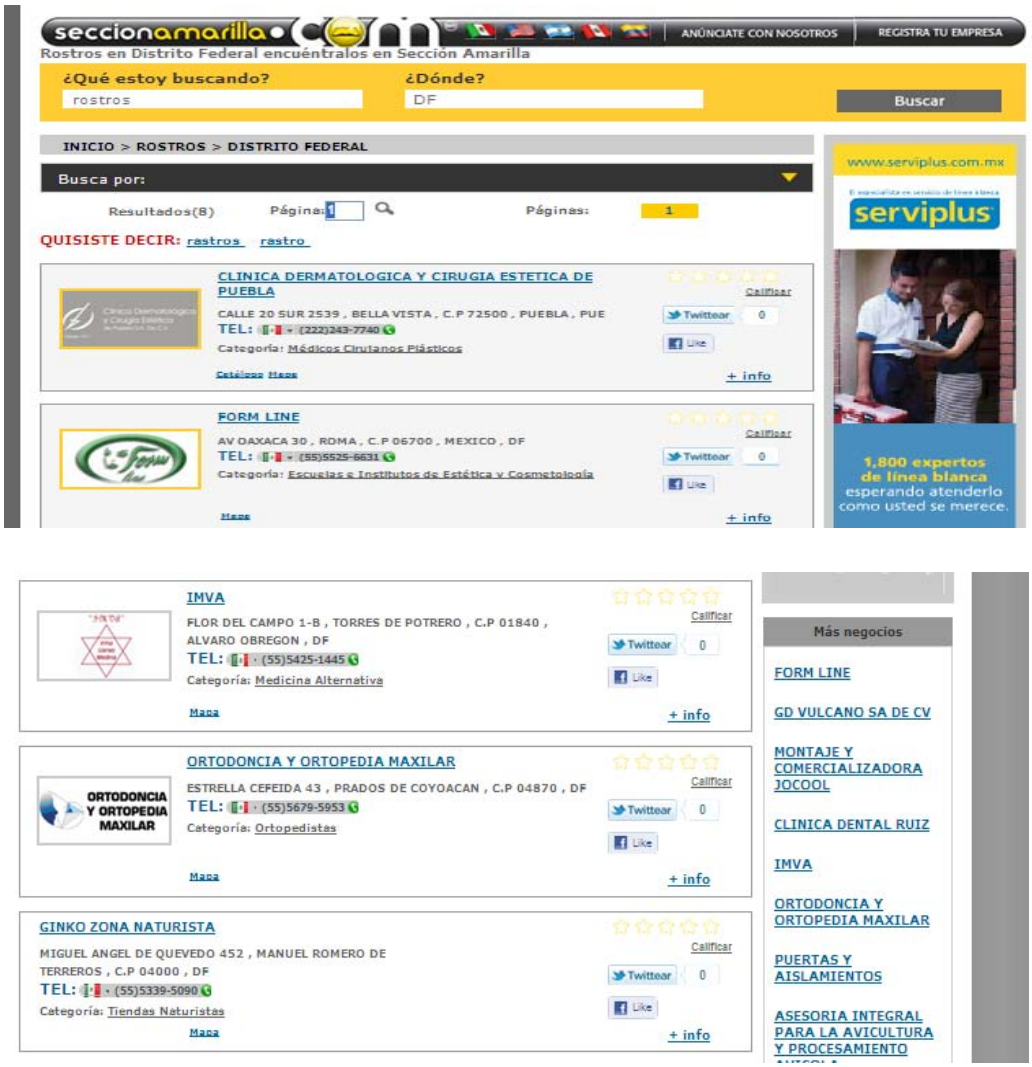
entenderá como la confusión generada a partir de una duplicidad de dos categorías, la cual conduce al usuario a interpretaciones equivocadas respecto a los contenidos del directorio; dicha confusión puede deberse también al uso indiscriminado de palabras clave y sinónimos no relacionados plenamente con la categoría en cuestión.(consúltese apéndices 1 y 2)

La duplicidad de categorías permite a los anunciantes publicarse en más de una al mismo tiempo. La elección de un criterio general (*escuelas de enseñanza artística*) o específico (*academias de canto*) sería de ayuda tanto para los anunciantes como para los usuarios. La propuesta es una revisión total de las categorías para encontrar las duplicidades y en todo caso, elegir la categoría más representativa para una comunidad o sector económico.

Respecto a la ambigüedad de las palabras clave y sinónimos, a diferencia del directorio impreso, los directorios *on line* deben integrar este tipo de palabras que funcionen como identificadores de una categoría. Lo anterior no siempre se aplica, puesto que no existe un estudio sociolingüístico que permita investigar cómo se conoce popularmente una categoría; por ejemplo, la inclusión de extranjerismos no está prevista (*jeans*, *pants*, *cabarets*, *body-piercing*).

La ambigüedad y la polisemia no son considerados problemas en estos directorios (por ejemplo, *clavo*, como alimento y *clavo* como herramienta), además de que el corrector ortográfico es de uso básico (*zapts* no devuelve una aproximación ortográfica a *zapatos*). Por ejemplo, al ingresar en la caja de búsqueda la palabra *rostros*, la página de resultados, como la que se puede ver en la figura 5, muestra categorías pertinentes (como *cosméticos, distribuidores, médicos cirujanos plásticos, escuelas* e *institutos de estética y cosmetología*) pero otros que poco pueden aportar a la búsqueda del usuario (*medicina alternativa, ortopedistas, tiendas naturistas*). ¿Qué puede ser lo que realmente quiere ver el usuario cuando escribe la palabra *rostro?* Quizá sólo quiera saber dónde hacen tratamientos faciales, o bien dónde comprar cosméticos. Pero de poco le sirve saber quién le vende a mayoreo, quién le puede distribuir, o saber que existe un ortopedista que le pueda ayudar con alguna cuestión sobre el rostro. De lo que se trata es de idear un sistema que pueda contemplar todas esas finas operaciones de búsqueda.

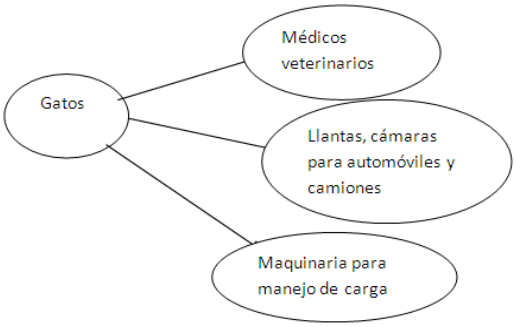
Figura 5. Representación de la búsqueda de la palabra clave *rostros*



La propuesta es una reorganización de los contenidos, ya sea por medio de categorías temáticas que permitan seccionar la información en grandes ramos del conocimiento o a través de categorías generales o específicas que no den lugar a la duda para saber qué puede encontrarse en ellas.

La resolución de la ambigüedad es un buen ejemplo de cómo funcionaría una organización temática. Actualmente, cuando se ingresa en el buscador la palabra *gato*, la cual presenta un caso de ambigüedad semántica, como se presenta en la figura 6, aparecen resultados de categorías referentes a herramientas y a médicos veterinarios; existe además una categoría que específicamente se llama *gatos hidráulicos y mecánicos*. Es comprensible que la palabra *gato* sea incluida en cada una de las categorías anteriormente señaladas, sin embargo ¿qué tan útil es la visualización de estos resultados si no existe una relación obvia de los mismos?

Figura 6. Ejemplificación de la ambigüedad semántica de la palabra *gatos*



El problema anterior podría resolverse mediante la división de la información en temas que desde la página principal pudieran seleccionarse. En lugar de escribir en el buscador la palabra clave, se tendría que seleccionar el tema del que se necesita obtener información. No habría por tanto resultados incoherentes ya que no se seleccionaron los que pudieran arrojar información no pertinente. Si bien esta solución puede dar resultados más precisos, tiene como desventaja el hecho de no ser rápida e instantánea: pasar por filtros no es lo más óptimo cuando se necesita urgentemente un médico.

La solución podría venir de la mano de la estadística. Con un índice de clasificaciones que no dé lugar a la ambigüedad, podría analizarse cómo es que los usuarios de ciertas categorías encuentran lo que buscan. Así, por ejemplo para las categorías *médicos veterinarios, llantas, cámaras para automóviles y camiones, maquinaria para manejo de carga* la palabra *gato* sería pertinente; pero entre las categorías *tubos de plástico, lavadoras a presión para limpieza automotriz, industrial y agrícola* estaría de más. El análisis estadístico ayudaría para saber cuál de las tres categorías pertinentes deber ser la primera en aparecer en la página de resultados.

Como se apreciará, la inclusión de herramientas lingüísticas como listas de sinónimos o palabras clave solucionaría problemas tan sencillos como el retorno de resultados relevantes. En el capítulo 3 se profundizará en ejemplos y propuestas sobre las problemáticas reseñadas en este apartado.

Capítulo 2: Lengua e Internet

La uniformización del mundo
empieza con la estandarización de
la lengua que usamos para designarlo[17]

[17] Mattelart, A. 2007.

2.1 Antecedentes y perspectivas de la sociedad de la información

Es común escuchar opiniones acerca de cómo las nuevas tecnologías han cambiado la interacción social. La disposición de recursos y documentos, y su libre acceso a ellos han contribuido al éxito de Internet y la web. A esta nueva sociedad, caracterizada por *el uso masivo de las tecnologías de la información y comunicación (TIC) para difundir el conocimiento y los intercambios en una sociedad*[18] se le ha llamado *sociedad de la Información*.[19]

Armand Mattelart explica en el libro *Historia de la sociedad de la información* que esta sociedad tiene un origen muy anterior,; entre los siglos XVII y XVIII, a partir de la entronización de la matemática como método infalible de razonamiento, la utopía de una sociedad con una lengua universal se materializa con la automatización de la razón a través del *calculus ratiocinator* o máquina aritmética. Leibniz comenta al respecto que*…el lenguaje de signos es el único que puede resolver las imperfecciones de las lenguas naturales que son otras tantas fuentes de discordia y de obstáculos para la comunicación.*[20]

A partir del sentimiento cientificista de la época, el pensamiento religioso y sus concepciones fueron reemplazados por una nueva organización del conocimiento formal, lógico y matemático, libre de ideologías.

Descartes fue otro filósofo que contribuyó a esta idea; en su carta a Mersenne[21] plantea ideas semejantes a las concebidas por Leibniz en relación con el problema que implica la lengua: *Al contrario, casi todas nuestras palabras tienen significados confusos; y el espíritu de los hombres está tan acostumbrado a ellas que esto le causa que no entienda casi nada perfectamente"*[22]*.*

En la carta expone las dificultades sociales a las que se enfrenta el hombre a partir de la lengua que usa; al mismo tiempo, detalla el proyecto para desarrollar una nueva lengua, libre de irregularidades, entendible para todos los hombres: *Inventar (hallar) una escritura, pues si vuestro hombre puede poner en su diccionario un solo símbolo que corresponda con aimer, amare, Φιλειν, y sus sinónimos, el libro escrito con esos caracteres podrá ser interpretado por todos aquellos que posean este diccionario.*[23]

---

[18] Téllez Valdez. *Derecho informático 3ª ed*. México: McGraw-Hill Interamericana Editores, 2004.
[19] Yus Ramos, F. *Ciberpragmática 2.0: nuevos usos del lenguaje en Internet*. Madrid: Ariel, 2010.
[20] Mattelart, A. 2007.
[21] Adam, Ch., y Tannery P. *Ouvres de Descartes*. Paris: J. Vrin, 1974.
[22] Adam, Ch., 1974.
[23] *Ídem*

Para iniciar este proyecto se necesitaba construir una base de conocimiento que estandarizará los saberes humanos y no dejara lugar a la confusión. Al respecto, en 1668 John Wilkins contribuyó al desarrolló de un lenguaje analítico que dividía los saberes del mundo en cuarenta categorías, subdivisibles entre sí. Esta curiosa taxonomía motivó, siglos después, a Jorge Luis Borges[24] a escribir sobre la imposibilidad de resolver el caos de los saberes humanos. Luego, Michel Foucault situó en su ensayo *Las palabras y las cosas*[25] el momento histórico del rompimiento entre las palabras y las cosas que representan, las cosas del mundo. Así, las palabras (y el saber mismo) se reorganizan en taxonomías *sin hogar ni lugar*.

Mattelart refiere que será hasta finales del siglo XIX que Paul Otlet y Henri LaFontaine[26] formalicen la idea de Wilkins, es decir, la organización del conocimiento a través de la sistematización de la información. Otlet tenía como objetivo hacer del conocimiento una red social, universal y técnica, accesible a todos. Este socialista belga consideraba que la diferencia de clases se originaba a partir de que la información sólo era accesible para una élite. Lo anterior lo impulso a desarrollar una metodología que permitiera la disponibilidad del conocimiento para todo aquel que lo necesitara. Junto con LaFontaine creó el *Repertorio bibliográfico universal* (RBU) para el Instituto Bibliográfico de Bruselas[27], el primer sistema de fichas bibliográficas conocido. El repertorio da pautas para generar instrumentos fundamentales para la gestión de conocimiento científico, basando toda su estrategia en la organización (clasificación) de los contenedores (documentos) donde se guarda el conocimiento representado, es decir, la información. Su propósito era más ambicioso pues imaginaba un proyecto similar en cada ciudad importante del mundo. En 1900, Otlet desarrolló junto con el ingeniero Robert Goldschmidt el almacenamiento de datos en micro-fotografía (microfilms).

Otlet se refirió a *una red universal de información y documentación*. Con esto, se inició conceptualmente la idea del mundo universal, donde las fronteras y sociedades dejan de existir, donde las redes han formado una civilización única. [28]

2.2 Ingeniería lingüística

En la actualidad, la sociedad ha integrado tan bien a su estilo de vida la tecnología y los servicios derivados de ella que pocas veces se llega a pensar en la gente que hace eso posible. En este punto, el uso de

[24] Borges, Jorge Luis, "El idioma analítico de John Wilkins", en *Otras Inquisiciones.* Madrid: Alianza Editorial, 1997.
[25] Focault, Michel. *Las palabras y las cosas*. México: Siglo XXI, 1967.
[26] Mattelart, A., 2007
[27] *Ídem*
[28] Platón definió los límites del tamaño de una ciudad como el número de personas que podían oír la voz de un solo orador. Hoy esos límites no definen una ciudad sino una civilización" *ídem* p. 56

traductores automáticos, correctores ortográficos y herramientas para el reconocimiento de voz son resultado del trabajo de profesionales de la lengua.. Al respecto, los lingüistas han encontrado en el desarrollo tecnológico de la sociedad de la información un panorama profesional distinto al que habitualmente podrían elegir.

Las oportunidades que las industrias de la lengua ofrecen a lingüistas exigen un perfil profesional más completo: aunado al conocimiento lingüístico, es necesario la adquisición de conocimientos informáticos, además de una fuerte capacidad de abstracción, razonamiento lógico, capacidad de organización y estructuración de datos.

Las interacciones comunicativas de esta sociedad tecnológica se han complementado con las llamadas *tecnologías de la comunicación* que tienen su apoyo en el uso de sistemas informáticos y de datos lingüísticos formales. El procesamiento de materiales escritos y sonoros constituye la materia sobre la que giran las investigaciones sobre tecnologías lingüísticas. A grandes rasgos, la ingeniería lingüística se centra en el tratamiento computacional del lenguaje natural y en su aplicación para la solución de problemas de comunicación tecnológica. A continuación, se revisará la definición y las perspectivas a futuro que tiene esta nueva área multidisciplinar.

2.2.1 Definición

La ingeniería lingüística se define como *la aplicación de los conocimientos sobre la lengua al desarrollo de sistemas informáticos que puedan reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas*[29]. Esta disciplina es una herramienta de apoyo para mejorar la utilización de la lengua en los sistemas computacionales, al asimilar, analizar, seleccionar y presentar la información para que las máquinas lleguen a *entender* el lenguaje natural y puedan satisfacer las necesidades de información de los usuarios de Internet. El uso de la lengua como herramienta tecnológica contribuirá a mostrar resultados con mayor precisión, lo cual permitirá superar el problema de exceso de información. Se trata, por tanto, de llegar a disponer de datos suficientes para procesar la información lingüística en función de las necesidades de la aplicación (por ejemplo, los sistemas de traducción, los correctores ortográficos automáticos, los lematizadores, entre otros).

La comprensión del lenguaje requiere de un análisis que puede referirse al estudio de las palabras (analizadores morfológicos), de las frases (analizadores sintácticos), del significado (analizador semántico), o de las cualidades comunicativas del texto (analizador pragmático). Los pasos básicos de la ingeniería lingüística consisten en la introducción del material en un sistema informático proveniente de distintas fuentes ya sean documentos sonoros o escritos. Sus fines son:

- Reconocer la lengua en la que está elaborado el material.
- Comprender el significado (entendido como el proceso para relacionar léxica, semántica y pragmáticamente los contenidos de acuerdo a criterios ontológicos de organización del conocimiento) del material introducido. En este punto, es conveniente aclarar que la web tiene un significado intrínseco para los usuarios, esto es, significa por la misma capacidad humana de dotar a los elementos del mundo de sentido. Sin embargo, como se ha mencionado, la web sólo es un repositorio de documentos y fuentes de referencia que carece de significado o de relación. Al final, sólo se trata de líneas de código trabajando.
- Interpretar el nivel lingüístico particular (morfológico, semántico o pragmático).
- Generación de interfaces para presentar los resultados obtenidos.

2.3 Representación del conocimiento y lingüística cognitiva

[29] Definición del informe para la Comisión Europea del IV Programa Marco para el apoyo de las actividades de investigación y desarrollo tecnológico dentro del sector de la ingeniería lingüística.

¿Qué es la comprensión? La respuesta a esta pregunta es fundamental para el desarrollo del Procesamiento de lenguaje natural. Se entiende por comprensión al proceso cognitivo que integra correctamente un nuevo conocimiento a los ya existentes en un individuo. Comprender es transformar ese conocimiento en una representación útil para lograr un objetivo determinado. Con base en lo anterior, ¿para qué se necesita que una máquina (computadora) entienda el lenguaje humano? En el paradigma tecnológico actual el desarrollo de sistemas informáticos que puedan comunicarse directamente con los humanos a través del lenguaje usado habitualmente, sin necesidad de otros medios gráficos (como las interfaces) es una prioridad. El objetivo es integrar la comprensión humana (lo que incluye las emociones y los matices de la lengua) a las máquinas.

2.3.1 Lingüística cognitiva

Como primer paso para lograr lo anterior es necesaria la integración de las herramientas que permitan la comprensión de un texto tal como un humano lo haría. El desarrollo de analizadores de los distintos niveles de la lengua (morfológico, fonológico, sintáctico, semántico, pragmático) es útil para tal fin. En los siguientes apartados se analizará cada una de estas herramientas lingüísticas.

El estudio de la lengua se ha enriquecido a partir de la consideración de los procesos cognitivos por medio de los cuales se origina el lenguaje. Esta perspectiva considera a los modelos de percepción y de memoria como factores determinantes para el lenguaje. María Josep Cuenca explica en su libro *Introducción a la lingüística cognitiva*[30] que la lingüística y la ciencia cognitiva tienen su punto de convergencia a mediados del siglo pasado. En 1948, el Instituto Tecnológico de California organizó un simposio cuyo fin era comprender los mecanismos mentales relacionados con el aprendizaje. Sin embargo, la investigación en el campo de la cognición no prosperó debido en parte a la escasa comprensión de la mente humana y a la corriente de la psicología conductista, predominante en aquellos días. A finales de los años 50 se retomó el interés de los investigadores, principalmente por las teorías generativistas de Noam Chomsky, lingüista estadounidense.

Como se ha dicho, la lingüística cognitiva relaciona el conocimiento lingüístico del hablante con otros procesos cognitivos, como la percepción, la memoria o la atención. Algunos autores proponen como postura filosófica al experiencialismo, que sostiene que *el procesamiento cognitivo depende de la estructura global del sistema conceptual y no simplemente de operaciones aisladas y símbolos, por lo tanto el sistema conceptual se basa en la percepción y la experiencia física y social. El experiencialismo analiza al lenguaje*

[30] Cuenca, M. J., y Hilferthy, J. *Introducción a la lingüística cognitiva*. Barcelona: Ariel, 1999.

como una experiencia corpórea[31], esto es que está relacionado con la experiencia física del hombre. Esta disciplina propone *una teoría unitaria de la mente, una estructura común para todos los procesos mentales relacionados con el conocimiento*. Se reconocen tres líneas de estudio: la teoría de prototipos, la semántica cognitiva y la teoría de la metáfora. La lingüística cognitiva ha puesto interés en problemas relativos a la categorización y a los modos en los que se relaciona el conocimiento lingüístico con el conocimiento del mundo.

2.3.2 Representación del conocimiento

Lidia Cámara de la Fuente comenta en un artículo publicado en la web *La representación lingüística del conocimiento y su relevancia en la ingeniería lingüística*[32] que la organización del conocimiento es una disciplina de origen reciente que analiza las leyes, los principios y los procedimientos para estructurar el conocimiento especializado y establecer el diseño en el que se basa el conocimiento de cualquier dominio. Aristóteles estudió esta materia, en su empeño por fragmentar el mundo para lograr una mejor comprensión del mismo y clasificarlo. Desde entonces, los cambios que ha sufrido esta disciplina van desde concepciones metafísicas y especulativas hasta las que tienen lugar en la actualidad, como las concepciones físicas, informático-pragmáticas, basadas en ciencias básicas como la lingüística, las ciencias del conocimiento y las ciencias de la información y la comunicación. Como se apreciará, el campo de investigación de la organización del conocimiento es multidisciplinar.

En el caso particular de la lingüística, debido a la influencia de las tecnologías de la información, la gestión de recursos lingüísticos se ha convertido en una necesidad para instituciones privadas y públicas. Se trata de una industria del conocimiento que contribuye a la creación de más conocimiento a partir del mismo. Es pues indispensable contar con métodos, herramientas y recursos lingüísticos.

Aplicaciones informáticas propias de la ingeniería lingüística, como la traducción automática, el reconocimiento de voz, la recuperación de información a través de Internet e Intranet, resúmenes automáticos, entre otros, necesitan trabajar con el conocimiento representado en forma de productos terminográficos, sistemas conceptuales y otros recursos lingüísticos.

Cámara de la Fuente refiere que desde una perspectiva cognitiva, los sistemas desarrollados en el área de la ingeniería lingüística están relacionados intrínsecamente a la ingeniería del conocimiento porque

---

[31] *Ídem*

[32] Cámara de la Fuente, Lidia. "La representación lingüística del conocimiento y su relevancia en la ingeniería lingüística" [en línea]. *Hipertext.net*, *núm. 2*, 2004 .http://www.hipertext.net/web/pag224.htm.

sus planteamientos se basan en estructuras lingüísticas que conforman sistemas conceptuales propios del hombre. Esta visión cognitiva se está aplicando en las investigaciones del PLN. Mientras que en los inicios de la ingeniería lingüística todo se basaba en estadísticas y la inclusión de analizadores morfológicos y sintácticos, ahora se da una relevancia importante a la semántica, resultado de los avances en la ciencia del conocimiento. Existe un cambio en el paradigma del conocimiento, de uno que era eminentemente lineal a otro que acepta complejidades en el sistema.

La ingeniería lingüística es una ciencia aplicada, resultado de la representación de los lenguajes artificiales, de deducciones lógicas a partir de la lingüística textual, la lingüística computacional, la terminología y la organización del conocimiento.

Las áreas de aplicación de la ingeniería lingüística son:

•Adquisición de conocimiento

-Identificación de unidades de conocimiento.[33]

-Extracción de información que representa unidades de conocimiento. El sistema será capaz de extraer información sobre un tema en específico.

•Modelado de conocimiento

-Desarrollo de metadatos interpretables de forma digital (inclusión de datos lingüísticos, como palabras clave y sinónimos, dentro del código de programación).

-Arquitectura del conocimiento (a través de taxonomías).

-Macroestructura: estructura para la clasificación, indización y recuperación del conocimiento (sistemas conceptuales) (creación de ontologías).

•Representación de conocimiento

-Creación de términos.

-Recopilación de unidades de conocimiento.

-Identificación de las relaciones semánticas entre unidades de conocimiento.

-Microestructura: estructuración de unidades o grupos de conocimiento (redes léxico- semánticas).

-Macroestructura: estructura para la clasificación, indización y recuperación del conocimiento (ontologías).

•Infraestructura para el desarrollo de ingeniería de conocimiento

-Bases de conocimiento.

[33] Se debe entender como unidad de conocimiento para el contexto de la web a la información existente en la misma. En este sentido, cada documento web es una unidad de conocimiento.

-Reglas de inferencia.

-Recuperación de conocimiento.

-Interacción y diseño de interfaces humano-máquina.

La representación del conocimiento tiene como objetivo simular los procesos a través de los cuales la mente humana organiza la información y aplicarlos a un sistema de inteligencia artificial[34]. Cualquier aplicación que utilice el conocimiento como clave para la resolución automatizada de problemas cognitivos deberá estructurar un proyecto de representación de conocimiento[35].

2.4 Procesamiento del lenguaje natural

¿Podrá una máquina comprender al hombre? Esta idea ha rondado la mente humana desde siempre; basta recordar aquellas caricaturas en las que las máquinas comprendían e interactuaban con los hombres; quizá el dicho la *ficción supera a la realidad* sea sólo un reto que los visionarios imponen a los hombres para superar los obstáculos de la sociedad. Sin embargo, para que este sueño se cristalice falta más investigación; no basta que la interacción con una computadora muestre comprensión, necesita demostrar razonamiento lógico, esto es, capacidad de pensar, decidir, actuar. Para lograr tal fin, es necesaria una comprensión de los procesos cognitivos por los que atraviesa el lenguaje en la mente humana. A esta nueva materia de estudio se le llama Procesamiento del lenguaje natural (PLN). Esta disciplina se ocupa de la comprensión y la interpretación del lenguaje natural en el marco de aplicaciones concretas. El PNL se encarga de gran parte de estos procesos, que echan mano de lematizadores, analizadores morfológicos, analizadores sintácticos y modelos semánticos.

2.4.1 Historia, definición y objetivos

A finales de los años cuarenta y como parte del espionaje entre Estados Unidos y Rusia durante la Guerra Fría, se inició la investigación entorno a la traducción automática (TA). Sin embargo, estos esfuerzos fracasaron debido a la poca capacidad tecnológica, pero también por la falta de estudios lingüísticos que comprendieran el tema de manera satisfactoria. Con la aparición de la TA se inicia el desarrollo de la inteligencia artificial (IA). De esta última se desprende el PLN que se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio de lenguajes naturales. Gran parte de las aplicaciones del PLN son implementables en la traducción

[34] La inteligencia artificial es una disciplina dedicada a la construcción de máquinas que implementan tareas propias de los humanos dotados de inteligencia (pensar, evaluar, actuar).

[35] *ídem*

automática, la indización automática de textos, la interacción hombre-ordenador en lenguaje natural, la elaboración automática de resúmenes y la extracción y recuperación de información.
Los objetivos del PLN son:

- Generación de interfaces en lenguaje natural: como ya se ha explicado, la interfaz es la *cara* de los programas informáticos. El objetivo es lograr que esas interfaces integren a los algoritmos de búsqueda tanto el sistema gráfico basado en íconos y botones, como la escritura y la voz.
- Procesamiento de datos: el exceso de información en la web ha propiciado que se piensen en nuevas formas para recuperar la información de forma precisa. Actualmente, muchos motores de búsqueda basan sus algoritmos de búsqueda y recuperación de datos a través de palabras clave; sin embargo, el uso de este tipo de palabras ha ocasionado un cierto grado de imprecisión en el proceso de recuperación de la información.
- Traducción automática: esta área de investigación es la que recibe más atención en cuanto a su desarrollo científico y tecnológico.

El PLN diseña mecanismos para comunicarse, que sean eficaces computacionalmente y simulen la comunicación humana. Estos procesos abarcan la comprensión del lenguaje, así como aspectos generales cognitivos y de organización de conocimiento en humanos.

Algunos de los retos que enfrenta el PLN se enfocan en resolver la ambigüedad propia de la lengua. En la práctica, este tipo de problemas se resuelve gracias al contexto de la oración. Sin embargo, para aplicarlo al PLN se necesitan recursos léxicos como gramáticas, diccionarios, bases de conocimiento, además de echar mano de otras disciplinas como la estadística. Los problemas de ambigüedad abarcan no sólo a la TA, sino también a la recuperación y extracción de la información. En este sentido, lo que se evitaría sería la recuperación de información irrelevante. En el caso particular de los directorios comerciales web los problemas de recuperación de información se deben principalmente a las palabras clave asignadas a una categoría con la cual no guarda relación intrínseca (por ejemplo la palabra *corazón* para la categoría *acero inoxidable, venta* de) y a la polisemia de una palabra; siguiendo con el ejemplo anterior, la palabra *corazón* bien puede asignarse a categorías relacionadas con la salud (*cardiólogos*), pero también a aquellas que tengan que ver con la maquinaria pesada (*motores eléctricos, venta,* en el caso de que alguna pieza de ese motor eléctrico lleve por nombre *corazones y ejes para baleros*) e incluso con la industria alimenticia (*alimentos congelados, venta de,* en el caso de que una palabra clave sea *corazones de alcachofas*).

Una solución para los casos referidos puede ser el uso de estadísticas que muestren la tendencia en el uso de una palabra en específico, esto es, analizar qué es lo que más buscan los usuarios cuando escriben

la palabra *corazón.* Así, esa palabra en particular sólo se podrá usar en contextos pertinentes (en el ejemplo, en aquellas relacionadas con el área de la salud) y se restringiría en aquellos entornos ambiguos.

Los problemas de ambigüedad no sólo atañen al nivel léxico-semántico sino también al sintáctico. El estudio de la ambigüedad referencial y estructural va enfocado a la gramática y el análisis semántico y pragmático de las oraciones en el desarrollo del texto. El PLN se ha encargado de trabajar sobre estas problemáticas y ha puesto en marcha proyectos que tienen bien delimitados sus contextos oracionales y semánticos (como puede ser el caso de las áreas técnicas). Como se apreciará, el campo de trabajo en el área de PLN es vasto y con amplias perspectivas a futuro.

2.4.2 Procesamiento lingüístico del lenguaje natural

El procesamiento se basa en la aplicación de técnicas y reglas para codificar el conocimiento lingüístico. Cada documento recuperado en la web es analizado a partir de diferentes niveles lingüísticos a través de herramientas lingüísticas que incorporan al texto anotaciones propias de cada nivel. Los pasos del análisis incluyen:

- Análisis morfológico: asignan a cada palabra su categoría gramatical a partir de rasgos morfológicos identificados.
- Análisis sintáctico: consiste en observar cómo se relacionan y combinan entre sí las palabras para formar sintagmas y frases. Se utilizan gramáticas, llamadas en el medio informático *parsers*, que son formalismos descriptivos del lenguaje y tienen por objetivo fijar la estructura sintáctica del texto.
- Análisis semántico: se trata de obtener el significado de las frases que lo componen. Es una representación semántica de las frases a partir de los elementos que las forman.

2.4.2.1 Herramientas lingüísticas[36]

Como ya se ha comentado, las herramientas lingüísticas son aquellas que ayudan a la interpretación de los datos como parte del PLN. Joaquim Listerri en su libro *Tecnologías del lenguaje*[37] describe que estas

[36] Listerri, J. "Lingüística y tecnologías del lenguaje" [en línea]. *Lynx, Panorámicas de Estudios Lingüísticos, vol. 2*, 2003 http://liceu.uab.es/~joaquim/publicacions/Llisterri_03_Linguistica_Tecnologias_Lenguaje.pdf

herramientas suelen dividirse entre los que se ocupan de la lengua escrita y los que tienen por objeto el habla. El desarrollo de estas tecnologías y sus aplicaciones requieren disponer de los llamados recursos lingüísticos, como diccionarios, gramáticas y corpus. En este trabajo sólo se estudiarán las herramientas lingüísticas propias de la lengua escrita (lo que no significa que en la red no existan recursos computacionales para analizar los documentos sonoros). Se dará prioridad a los aspectos lingüísticos sobre los tecnológicos; no se pretende presentar exhaustivamente un tema en específico, y más bien se dará una idea general y puntual sobre cada una de estas herramientas.

Estas herramientas, que no están visibles a los usuarios finales de la web, son imprescindibles para el desarrollo de las tecnologías lingüísticas. Se trata de programas que realizan las tareas que habitualmente un lingüista (o cualquier hablante de la lengua) haría: extraer la raíz de las palabras, segmentarlas en morfemas, asignarles categoría gramatical y determinar a qué parte de la oración pertenecen para indicar la función sintáctica de cada uno de los constituyentes que conforman la oración. En los siguientes apartados se describirán cada una de estas herramientas, mientras que en el capítulo siguiente se explicará la metodología para su desarrollo.

2.4.2.1.1 Analizadores morfológicos

Un analizador morfológico, también llamado *Part of Speech Tagger* o *POS Tagger*, analiza la forma y composición de las palabras (véase tabla 1: ejemplificación del análisis morfológico de la oración *Las tecnologías lingüísticas mejorarán las economías*). Como se ha mencionado, el primer paso para todo análisis lingüístico computacional parte del análisis morfológico. Es útil recordar que la morfología estudia la estructura interna de las palabras. Su unidad de trabajo son los morfemas, los cuales poseen significado léxico o gramatical.

En un sistema computacional, la imposibilidad para listar todas las palabras del diccionario con sus respectivas variantes ha hecho de esta herramienta una parte fundamental del PLN. Los objetivos de los analizadores morfológicos son:

- Identificar palabras pertenecientes a idiomas específicos.
- Reconocer y aislar los diferentes elementos que componen las palabras. Estos elementos son los lexemas (entendido éste como la unidad mínima que no presenta morfemas gramaticales, como por ejemplo *sol*)y afijos (prefijos y sufijos).

[37] Martí, M.A. *Tecnologías del lenguaje*. Barcelona: Editorial UOC, 2003.

- Obtener las entradas del diccionario (lemas) para relacionarlas con las formas léxicas del diccionario electrónico, las cuales incluyen todas las entradas que el sistema es capaz de reconocer y generar.
- Describir la información morfológica relevante de cada lema (número, género, tiempo verbal, persona, véase el ejemplo de la tabla 1 al respecto).
- Asignar a las palabras una o varias categorías gramaticales (nombre, verbo, adjetivo, determinantes, etc.) En caso de que una palabra pertenezca a más de una categoría gramatical se debe desambiguar atendiendo al contexto, por ejemplo la palabra *casas* podría ser un nombre o un verbo según el contexto en el que aparezca.
- Relacionar los elementos compositivos de las palabras a partir de reglas morfológicas.

Tabla 1. Ejemplificación del análisis morfológico de la oración *Las tecnologías lingüísticas mejorarán las economías*

Palabra	**Lema**	**Descripción morfológica**
Las	la	Artículo definido, género femenino, plural
tecnologías	tecnología	Nombre común, género femenino, plural
lingüísticas	lingüística	Adjetivo calificativo, género femenino, plural
mejorarán	mejorar	Verbo principal, modo indicativo, tiempo futuro, tercera persona, número plural
las	la	Artículo definido, género femenino, plural
economías	economía	Nombre común, género femenino, plural

El criterio para la identificación de una palabra es la localización de ésta entre espacios en blanco o entre un espacio en blanco y un signo de puntuación. Como se verá en la tabla 2, el primer paso es introducir una palabra para que el analizador identifique el lema al que corresponde, su categoría gramatical, género y número. Este análisis permite obtener las posibles lematizaciones de la palabra de acuerdo a la base de morfemas pre-identificados como tales en la base de datos léxica.

Tabla 2. Ejemplificación del análisis de la palabra *padres*

Interpretación 1	
Lema:	Padre
	Nombre común

	Género masculino
	Número plural

2.4.2.1.2 Lematizadores

Los lematizadores son programas que asocian una forma flexionada con su lema (la palabra o entrada que aparece en el diccionario). Esta herramienta analiza la palabra separando su raíz de los afijos flexivos y derivativos (sufijos como género, número y categoría gramatical) En la tabla 3 se analiza el proceso de lematización del lema *libro*. Lo primero es identificar su raíz o lema, para que pueda añadirse el morfema flexivo *–s* para formar *libros* o el morfema derivativo *–ito* para formar *librito.*

Tabla 3. Ejemplificación del proceso de lematización

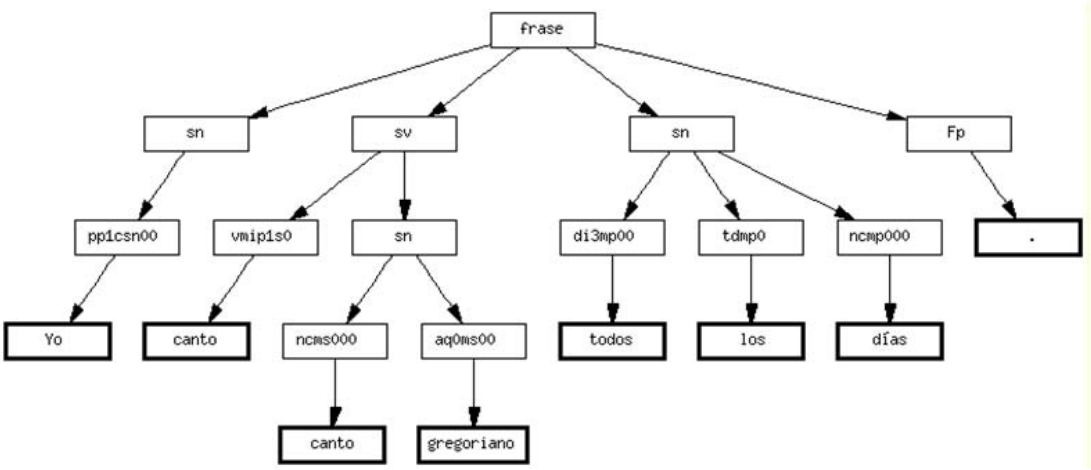
Lema	**Raíz**	**Sufijos**
Libro	Libr-	os
Libro	Libr-	ería
Libro	Libr-	eta
Libro	Libr-	ito
Libro	Libr-	eto

Los lematizadores sirven para definir patrones de corrección ortográfica, a partir reglas de formación de palabras que puedan detectar errores en todas las formas flexionadas de una misma palabra sin necesidad de que todas ellas estén incluidas en una base de datos léxicos. El desarrollo de esta herramienta contribuye a encontrar palabras que contengan una raíz determinada y no sólo las que coincidan con el término exacto que se ha indicado en la caja de búsqueda. Se recomienda el uso de lematizadores en el desarrollo de redes léxico-semánticas, como *WordNet*, que permiten localizar palabras relacionadas con el término de la búsqueda; de esta manera, al introducir la palabra *libro* también se obtienen resultados con palabas como *librería*, *libros*, *libretos* o *libreros*. En el capítulo siguiente se explicará las metodologías para el desarrollo de estas herramientas.

2.4.2.1.3 Analizadores sintácticos

Esta herramienta analiza automáticamente las dependencias estructurales entre las frases, o grupos de palabras, para determinar la función sintáctica que desempeña cada una de ellas. Al igual que el análisis morfológico, para identificar las frases se consideran factores tipográficos e informáticos, por ejemplo una cadena de palabras limitada a la izquierda por un espacio en blanco o un signo de puntuación, y a la derecha por otro signo de puntuación (en la figura 7 puede verse la representación del análisis de una oración realizado por un analizador sintáctico).

Figura 7. Representación del análisis de la oración *Yo canto canto gregoriano todos los días* realizador por un analizador sintáctico



Para las tres herramientas presentadas se requiere de una formalización del conocimiento lingüístico, ya que se trata de una emulación automática de las operaciones cognitivas llevadas a cabo por un lingüista. La presencia de este profesional es imprescindible en el desarrollo de este tipo de herramientas. Su trabajo podría dividirse en dos fases: en la primera de ellas consiste en definir las etiquetas o categorías morfológicas y léxicas que se asignarán al análisis, la segunda en validar y refinar el análisis hecho por las máquinas.

2.4.2.2 Recursos lingüísticos

Los recursos lingüísticos son esenciales para el desarrollo de las aplicaciones propias de las tecnologías del lenguaje. Abarcan tres categorías: corpus, diccionarios y gramáticas (sobre esta última se discute si pertenecen más bien a las herramientas lingüísticas).

2.4.2.2.1 Corpus y anotación

Un corpus se define como el conjunto de textos que constituyen una muestra sobre el uso de la lengua. Éste puede ser oral o escrito. Los del primer tipo recogen muestras de grabaciones sonoras y de transcripciones ortográficas de la lengua hablada. Los del segundo tipo se enfocan principalmente en la obtención de conocimiento lingüístico para la conversión de texto en habla, el entrenamiento de modelos acústicos para el reconocimiento de voz o el diseño de sistemas de diálogo. Los corpus escritos constituyen la materia prima sobre la que se analizará los datos lingüísticos. Este análisis incluye marcas que definen su estructura con anotaciones lingüísticas. Estas anotaciones son primordiales a nivel morfológico, sintáctico, semántico, pragmático o textual, utilizando para este fin las herramientas lingüísticas antes descritas. Se requiere de un corpus anotado manualmente para que un anotador automático pueda realizar eficazmente su tarea. Los lingüistas dedican mucho tiempo a ello ya que de esto dependen los resultados que se obtendrán. La ambigüedad es uno de los problemas a los que se enfrenta un lingüista al momento de la anotación. Ésta puede ser gramatical (una misma palabra desempeña funciones distintas, figura 8) o semántica (una misma palabra con idéntica función gramatical tiene significados diferentes, figura 9). Para la resolución de estas cuestiones se necesita una desambiguación manual, incorporada a medida que sean detectadas las excepciones.

Figura 8. Ambigüedad gramatical

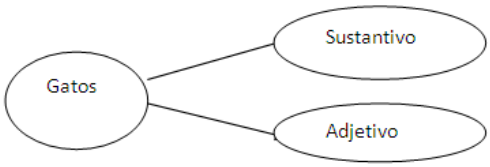
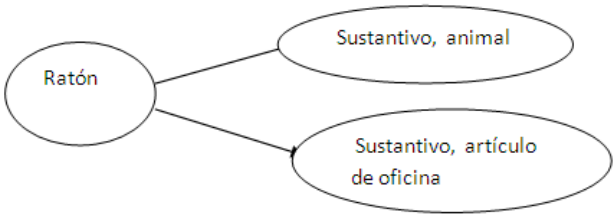


Figura 9. Ambigüedad semántica



Al igual que en la anotación morfológica, la anotación gramatical y semántica requiere de una revisión manual por parte de un lingüista para detectar excepciones y mejorar el sistema gradualmente. En cuanto a la anotación semántica y pragmática de un corpus, actualmente se desarrollan sistemas automáticos que exigen un conocimiento lingüístico necesario para definir las categorías (como sujeto, objeto directo, complemento circunstancial) y las etiquetas (sustantivo, verbo, adjetivo, entre otros) que se emplearán en la anotación.

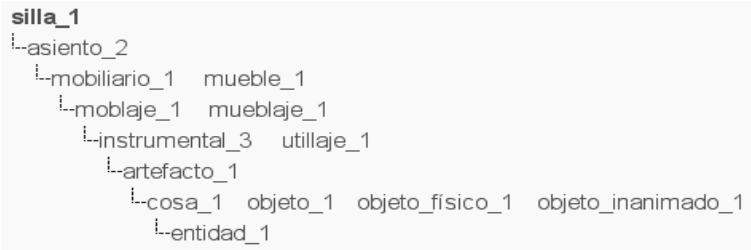
Para el propósito de este trabajo, se llamará *corpus* al índice de clasificaciones de *Sección amarilla*, a partir del cual se identificarán errores y generarán propuestas lingüísticas.

2.4.2.2.2 Recursos léxicos

Los recursos léxicos más utilizados para el desarrollo de herramientas lingüísticas son los léxicos computacionales y las redes léxico-semánticas. Un léxico computacional no es un diccionario tradicional; más que la definición de la palabra contiene información morfológica, sintáctica y semántica relevante para las diversas aplicaciones de procesamiento del lenguaje y para la incorporación a las herramientas del análisis automático. Las redes léxico-semánticas estructuran el vocabulario en función de relaciones semánticas entre palabras que toman como base conceptos propios de la semántica léxica como sinonimia, antonimia, polisemia, homonimia, hiperonima, hiponimia o meronimia[38]. El recurso más conocido y estudiado es *WordNet* (véase figura 10 en la cual se muestra el análisis componencial de la palabra *silla*), desarrollado en la Universidad de Princeton. Tomándolo como modelo, se han hecho versiones en lenguas distintas al inglés, como el italiano, francés, español, griego o alemán. A este recurso se le conoce como *EuroWordNet.*

Figura 10. Representación del modelo léxico-semántico de *WordNet*

[38] *Sinonimia*: relación semántica de identidad o semejanza de significados entre determinadas palabras; *Antonimia*: palabras que tienen un significado opuesto o contrario. *Polisemia*: se presenta cuando una misma palabra tiene varias acepciones. *Homonimia: relación de semejanza que presentan dos palabras con significado diferente. Hiponimia*: palabra que posee todos los rasgos semánticos de otra más general, pero añade en su definición rasgos particulares que la diferencian. *Hiperonimia*: término general que puede ser utilizado para referirse a la realidad nombrada por uno más particular. *Meronimia*: es la relación de partes que corresponde a un todo, por ejemplo una bicicleta está constituida por pedales, llantas, manubrios, cadena, asiento). Cf. Coseriu, Eugenio. *Introducción a la lingüística.* Madrid: Gredos, 1986.



Las redes léxico-semánticas constituyen un recurso utilizado en la anotación semántica de corpus y muestran un potencial interesante para la recuperación y extracción de información. Este tipo de recursos pueden considerarse como ontologías en el sentido que establecen una organización de conceptos. Se puede apreciar que los léxicos computacionales y las redes léxico-semánticas son formalizaciones del conocimiento léxico, morfosintáctico y semántico. Para que tengan la coherencia deseada es necesaria la intervención de un lingüista que supervise la eficacia de estos recursos.

2.4.2.2.3 Gramáticas

Una gramática computacional puede entenderse como una descripción formalizada del conocimiento lingüístico que, en el marco del procesamiento del lenguaje, puede ser empleada tanto por las herramientas de análisis como por las de generación de textos. Por esta razón se considera un recurso, junto a los corpus y a los léxicos, para el desarrollo de sistemas que realicen las operaciones que hemos descrito anteriormente.

Una gramática requiere un formalismo para expresar la información lingüística que contiene. Si en una primera etapa se utilizaron las gramáticas de estructura sintagmática –o gramáticas de estructura de frase-, el procesamiento sintáctico en los últimos años se ha centrado en las gramáticas de unificación y en la gramática de restricciones (*Constraint Grammar*). Las primeras reciben este nombre por el procedimiento que se aplica para combinar la información contenida en las categorías gramaticales, y tienen como principal característica la codificación de la máxima información posible en el léxico, al que se incorporan rasgos sintácticos y semánticos. Las gramáticas de restricciones parten de la anotación de las posibles funciones sintácticas de una palabra, para realizar después una desambiguación y seleccionar la función adecuada en una oración concreta.

Existen también aproximaciones, como la de la sintaxis léxica, que integran gramáticas y diccionarios electrónicos. Los diccionarios contienen, para cada forma, el lema a la que está asociada, la clase distribucional a la que pertenece y sus propiedades morfológicas. Las gramáticas consisten, en este marco teórico, en una formalización de las propiedades de los predicados que se encuentran en el diccionario.

No hace falta insistir de nuevo en que la construcción de una gramática computacional es una tarea imposible de abordar sin un detallado conocimiento de la estructura morfológica, sintáctica y semántica de la lengua. Cabe destacar también que los formalismos de unificación están teniendo una cierta repercusión en la teoría sintáctica como alternativas a otros modelos, con lo que se establece un diálogo entre los planteamientos propios de la lingüística computacional y los de la teoría lingüística.

Capítulo 3: Reporte de actividades profesionales

La empresa RedCúbica S.A de C.V fue fundada en el año 2003 por un grupo de inversionistas interesados en el naciente mercado electrónico. El objetivo, en un principio, fue el desarrollo de un directorio comercial *on line*, tipo *Sección amarilla*. La propuesta integraba además de una interfaz gráfica intuitiva para los usuarios, una revisión completa de sus categorías, la actualización de las mismas y el desarrollo de herramientas lingüísticas que pudieran facilitar la recuperación de información en ambiente Web. El proyecto evolucionó de ser un directorio comercial en línea a un motor de búsqueda, como *Google* o *Yahoo!*. Sin embargo, debido a las pérdidas registradas durante el periodo de preoperaciones, la empresa no logró su cometido y cerró en 2007 por falta de recursos financieros. El reinicio de operaciones no tiene fecha definida.

3.1 Metas laborales

Empecé a trabajar en la empresa RedCúbica S.A de C.V desde septiembre del año 2005 hasta julio del 2007. Desde un principio se plantearon los objetivos a realizar, así como un estimado de tiempo para alcanzarlos. Las metas fueron:

- Lanzar un directorio comercial web que subsanara las deficiencias que en ese entonces tenían proyectos similares.
- Integrar herramientas lingüísticas al algoritmo de búsqueda para lograr resultados más precisos y coherentes.
- Generar un modelo comercial que ubicara cada negocio, servicio o producto comercializado en la ciudad.[39] Este punto es importante ya que a través de un *censo* se conocería con precisión el nombre que la costumbre le da a las *cosas* del mundo. De esta manera una categoría podría asociarse al uso popular o tradicional, lo que daría más posibilidades a los usuarios para encontrar información.

Los objetivos específicos del equipo lingüístico fueron:

- Analizar los problemas que enfrentaba el directorio web de *Sección amarilla* (véase apéndice 2).
- Analizar las categorías que integran el índice de clasificaciones, determinar problemáticas a partir de ellas y hacer una propuesta metodológica (teórica y práctica) para la creación de nuevas categorías.
- Analizar las deficiencias lingüísticas de los directorios comerciales en la web y en general de los buscadores usados en ese entonces, y proponer un sistema lingüístico integral.

Estos objetivos fueron alcanzados en seis meses a partir de septiembre de 2005. Las propuestas presentadas fueron las siguientes:

[39] Debido al secreto industrial, no se dan más detalles al respecto

- Creación de nuevas categorías o reactualización de las ya existentes a partir de factores sociales (costumbre, popularidad, demanda comercial).El criterio para crear nuevas categorías se verá en apartados posteriores.
- Creación de herramientas lingüísticas pertinentes para hacer más precisa la búsqueda de información; las herramientas propuestas fueron un lematizador, un analizador morfológico, un analizador semántico que considerara un algoritmo para casos de ambigüedad[40], además de la inclusión de recursos léxicos como listas de sinónimos, antónimos, extranjerismos y de redes léxico-semánticas que estructurarán las relaciones entre las palabras. Más adelante se profundizará en este punto.

Se proyectó alcanzar estas metas en máximo un año a partir de febrero de 2006; para cumplirlas fue necesaria la contratación de más lingüistas entusiasmados en el sector de las nuevas tecnologías. Debido al interés mostrado de mi parte, me fue encomendado el papel de *guía* para los lingüistas que ingresaban a la empresa. El departamento de análisis lingüístico procuró tener entre sus filas a tres profesionales del área, sin embargo, aunque muchos estaban interesados, el desconocimiento del tema fue un factor importante para la rotación de empleados. Lo anterior implicó la demora para alcanzar los objetivos.

A continuación se describirán las funciones y la metodología implementada para el desarrollo de los objetivos laborales.

3.2 Descripción de actividades y responsabilidades

3.2.1 Perfil

La formación de un profesionista recién egresado de la licenciatura en Lengua y Literaturas Hispánicas de la Facultad de Filosofía y Letras de la UNAM no contempla su incursión dentro del área tecnológica. Lo anterior se debe a la falta de materias relativas al tema dentro del plan de estudios. Los avances tecnológicos requieren profesionales de la lengua con una formación multidisciplinaria. En este sentido, un lingüista deberá estar dispuesto a compartir sus conocimientos sobre la lengua, y además estar abierto a aprender nuevas propuestas metodológicas que no siempre correponderán a la teoría aprendida, para lo cual deberá ser creativo a fin de poder plantear la mejor solución a las problemáticas que enfrenta la búsqueda de información en Internet.Aunado a lo anterior, deberá comprender la lógica del trabajo que tienen losequipos de programadores e ingenieros, para quienes la lengua sólo representa una serie de signos que

[40] Debido al secreto industrial, este punto sólo será descrito de manera superficial

hay que insertar dentro de algún punto de las líneas del código de programación. . En este punto, es necesario buscar soluciones que beneficien a ambas áreas.

3.2.2 Actividades

Las actividades realizadas en RedCúbica fueron:

- Búsqueda de referencias bibliográficas que sirvieran de marco teórico para la documentación de la patente; en este sentido, sería óptimo que en las facultades que imparten la licenciatura de Lengua y Literaturas Hispánicas se ofrecieran clases sobre la relación existente entre la lingüística y la tecnología, sobre todo por la importancia que a últimas fechas ha tomado el uso del lenguaje en el desarrollo del área tecnológica
- Análisis de los índices de clasificación de diversos directorios comerciales en línea (*Páginas útiles* de *Yahoo!*, *Sección amarilla* y *Kompass*, entre otros de menor relevancia comercial pero de igual modelo que los anteriores)
- Propuesta de nuevas metodologías para la creación de categorías
- Propuesta de recursos y herramientas lingüísticas pertinentes para el desarrollo del proyecto

Por ser la lingüista que mayor tiempo dedicó al estudio del tema tecnológico, me fue confiada la responsabilidad de mantener la calidad y las expectativas del trabajo, proyectar tiempos, supervisar a los otros integrantes del equipo, además de estar en constante comunicación con el equipo de desarrollo, integrado por el arquitecto web (quien se encargó de diseñar la interfaz) y los programadores. A continuación, se describirán de manera detallada cada una de las actividades realizadas en el periodo de trabajo.

3.3 Análisis y elaboración de categorías

3.3.1 Antecedentes teóricos

Dice Lakoff en su libro *Women, fire and dangerous things* que el proceso de categorización es fundamental para el pensamiento humano: *Every time we see something as a kind of a thing, for example a tree, we are categorizing.*[41]

Durante más de dos mil años, la teoría aristotélica ha impuesto su percepción sobre los modelos de categorización en la sociedad occidental. Este paradigma se ha mantenido a partir de la especulación *a priori* sobre la experiencia de la vida: *Over the centuries it simply became part of the background assumptions taken for granted in most scholarly disciplines*.[42]

¿Qué es una categoría? La teoría clásica la define como un concepto abstracto a través del cual se puede reconocer, diferenciar y clasificar las distintas entidades que conforman el mundo. Mediante la formación de categorías, el hombre puede conocer a mayor profundidad la realidad circundante. Formalmente, una categoría es una clasificación jerárquica de las entidades de la realidad. La regla de la categorización según la teoría clásica dice que entidades muy parecidas y con características comunes formarán categorías, mismas que pueden ser parte de categorías superiores si y sólo si compartieran características comunes entre ellas.

A partir del siglo XX, la teoría clásica ha sido fuente de debates, lo que ha propiciado la polémica sobre la cognición y los procesos para categorizar. Para muchos, la lógica de las categorías del modelo aristotélico es sencilla: *They were assumed to be abstracts containers, with things either inside or outside the category. Things were assumed to be in the same category if and only if they had certain properties in common. And the properties they had in common were taken as defining the category*.[43]

Sin embargo, teóricos como Wittgenstein o Eleanor Rosch convirtieron el problema de la categorización en un tema central para entender el proceso cognitivo de la mente humana. Wittgenstein, en su libro *Investigaciones filosóficas*[44] plantea el concepto de *semejanza de familia* a partir de conceptos que

---

[41] Lakoff, G. *Women, fire, and dangerous things, what categories reveal about the mind*. Chicago: The University of Chicago Press, 1990, pág. 8
[42] *Idem*
[43] *Idem*
[44] Wittgenstein. L. *Investigaciones filosóficas*. México: Instituto de Investigaciones Filosóficas, UNAM, 1988.

comparten uno o más rasgos en común. Se trata de conceptos que se parecen entre sí, aunque algunos de ellos tengan más rasgos en común (típicos) que otros (atípicos).

Por su parte, Eleanor Rosch analiza en su libro *Natural categories*[45] el campo de la cognición en relación con las categorías; en la década de los 70 realizó un trabajo sobre la categorización de los colores en hablantes de distintas lenguas. Concluyó que éstos se organizan en torno a uno básico, fácilmente diferenciable. A partir de lo anterior, comprobó que en el proceso de percepción de los colores era importante la percepción de los *focos cromáticos* a los que llamó *prototipos*; esto significó que no todos los ejemplares que un sujeto agrupa en una misma categoría resultan *buenos ejemplos* de la misma, lo cual demostraba la existencia de miembros más prototípicos que otros e imposibilitaba la definición de una categoría por condiciones necesarias y suficientes, tal como se venía haciendo con la teoría clásica. Rosch definió como *prototipo* al ejemplar más representativo de una categoría, el que más rasgos comparte con el resto de los miembros de esta y menos con los de otras categorías. Planteó las categorías como clases con límites difusos (a diferencia de la teoría clásica en la que las clases tenían límites definidos) en los que se encontrarían miembros limítrofes con categorías vecinas; de esta forma, la transición de estas categorías sería gradual.

Luis Fernando Lara explica en su libro *Curso de lexicología*[46] que la teoría de los prototipos se ha criticado por la falta de características propias de muchas de las categorías, lo que genera confusión, (por ejemplo, la categorización de la ballena como pez o como mamífero). Ante críticas como las mencionadas, George Lakoff planteó una nueva concepción de la teoría de los prototipos. Propone un modelo cognitivo idealizado compuesto por categorías radiales y asociaciones mentales Como se aprecia, este modelo toma en consideración los fenómenos de la metáfora y la metonimia(por ejemplo, un ave y un avión son unidos gracias a que comparten el vuelo como una característica común).. Lo anterior refuerza la idea de que el proceso de categorización es crucial para entender la mente humana. Lakoff afirma que la categorización es esencialmente un proceso que involucra no sólo la imaginación sino también la experiencia social: *…human categorization is essentially a matter of both human experience and imagination –of perception, motor activity, and culture on the one hand, and of metaphor, metonymy, and mental imagery on the other*.[47]

3.3.2 Metodología para la creación de categorías

[45] Rosch, Eleanor. "Natural categories". *Cognitive Psychology* vol. 4, no. 3 (1976): 328-350 p.
[46] Lara, Luis Fernando. *Curso de lexicología*. México: El Colegio de México, 2006.

[47] Lakoff, G., 1990

Para la creación de las categorías del directorio de RedCúbica S.A. de C.V. fue necesaria una base teórica sobre la cual pudiera definirse el modelo metodológico a seguir. Si bien, las categorías seguirían el modelo prototípico, se pensó que no era fácil romper con la tradición de la teoría aristotélica. Se decidió conservar la estructura del modelo clásico, adaptándolo gradualmente al prototípico. Así, por ejemplo, se conservaría la categoría *hoteles,* pero se debería hacer una distinción entre *moteles, hostales* o *casas de huéspedes*. Si alguna de estas categorías tuviera más subdivisiones se consideraría hacer la distinción pertinente. Sin embargo, privaría más un criterio medio que no creara confusión por ser demasiado general o demasiado específico.

3.3.2.1 Identificación de problemáticas

Como primer paso en la creación del índice de clasificaciones de RedCúbica S.A de C.V, se identificaron los problemas de los directorios comerciales en línea (ver apartado 1.3 Directorios comerciales web y apéndice 2 ubicado al final del texto). Se eligieron tres directorios para analizar su índice de clasificaciones; estos fueron la *Sección amarilla*, *Páginas útiles de Yahoo!* y *Kompass*, directorio comercial español. Básicamente, estos tres directorios tienen la misma información, aunque organizada de diferente manera. Las siguientes tablas muestran una comparación entre la información que incluye cada uno de ellos:

- Caso 1: identificación de categorías que pueden crear ambigüedad al no detallar lo que se hace con los productos que le dan su nombre. Esta ambigüedad es ocasionada por el uso de un criterio demasiado general o muy específico:

Tabla 4. Ejemplo de categorías ambiguas como resultado de criterios generales y específicos

Directorios	**Ejemplos**
Sección amarilla	Leche
Páginas útiles, *Yahoo!*	Quesos
Kompass	Puertas de madera con plancha de plomo para salas de radiología

- Caso 2: identificación de categorías con nombres diferentes, pero que ofrecen el mismo tipo de servicios o productos. A este tipo de categorías se le llamará categoría sinónimas puesto, al final de cuentas, se refieren a lo mismo. También se identificaron las categorías meronímicas, es decir,

derivadas de ellas. Estos dos tipos de categorías generan confusión en el usuario que consulta el directorio *¿qué categoría es mejor? ¿en cuál de las dos podrá encontrar lo que busca?* :

Tabla 5. Ejemplo de categorías sinónimas y meronímicas	
Directorios	**Ejemplos**
Sección amarilla	- Agencias de colocaciones -Agencias de selección de personal -Agencias de cobranzas, véase *cobranzas*
Páginas útiles, *Yahoo!*	- Clínicas de belleza, véase también *estética, clínicas de* - Salones de belleza -Uñas postizas
Kompass	-Sociedades de inversiones -Fondos de inversión

- Caso 3: Categorías vacías con nombres similares, una de ellas sólo sirve de referencia para la otra, es decir, no tiene información de los anunciantes:

Tabla 6. Ejemplo de categorías vacías referenciales	
Directorio	**Ejemplo**
Sección amarilla	Legaciones, véase *embajadas, consulados y legaciones*
Páginas útiles, *Yahoo*!	Fiestas, alquiler de equipos, véase *fiestas, artículos para*
Kompass	No tiene este tipo de categorías

Además, se identificaron problemas a partir del tipo de categoría:

- Profesionistas

Tabla 7. Ejemplo de categorías referentes a profesionistas		
Ejemplo	**Problemática**	**Solución**
-Médicos -Abogados -Maestros	Estas categorías están relacionadas dentro del índice de clasificaciones con las que designan lugares (hospitales, consultorios médicos, bufetes de abogados, escuelas)	En las categorías referentes a los profesionistas que ejercen de manera individual su actividad económica se debería comprobar que efectivamente realizaran la actividad dentro de la cual se han clasificado. De esta manera se evitaría que los anunciantes se clasificaran en categorías similares

- Lugares o instalaciones

Tabla 8. Ejemplo de categorías referentes a lugares o instalaciones		
Ejemplo	**Problemática**	**Solución**
-Talleres automotrices -Hospitales -Escuelas	Crean confusión con aquellas que designan profesionales.	En estas categorías sólo podrían clasificarse aquellos anunciantes que comprobarán la existencia de las instalaciones físicas en donde llevan a cabo sus actividades comerciales
-Misceláneas -Tiendas de conveniencia	No siempre se usa como nombre de una categoría la forma más común de llamarla	Se deberá hacer un censo que corrobore el nombre que la costumbre le ha dado a ciertas categorías (por ejemplo *tienditas, tienda de*

		la esquina)

- Productos

Tabla 9. Ejemplo de categorías referentes a productos		
Ejemplo	**Problemática**	**Solución**
-Productos de belleza -Artículos de belleza	Categorías duplicadas por el uso de palabras ambiguas dentro de su nombre	Se creó un criterio editorial para elegir entre palabras conflictivas, como *producto* y *artículo*.
-Broches de presión – fábricas	Lugares que designan la fabricación específica de cierto producto	Se crearían categorías específicas solamente si existiera un nicho de mercado importante, en caso contrario se aplicaría un criterio medio. Se aseguraría que las categorías del directorio de RedCúbica S.A de C.V fueran relevantes comercialmente
-Zapatos ortopédicos	Productos que podrían ser parte de un catálogo	Sólo serían nombre de categoría aquellos productos que por su importancia constituyeran por sí mismos un fuerte nicho de mercado; en este caso, el producto *zapatos ortopédicos* sólo sería parte del catálogo de la categoría *Productos ortopédicos*
-Cintas y discos para máquinas computadoras	Categorías sin actualizar el nombre por el que las	El nombre de la categoría debe ser acorde con lo que

	conocen; en el ejemplo se lee *máquinas computadoras* cuando se llanamente se denominan *computadoras*	se sabe de ese nicho de mercado. En este caso, para aquellos usuarios no especializados en el tema, el nombre no es el más indicado para identificar a la categoría (especificidad). Sería más fácil asociar estos productos con una categoría general como *hardware, venta de* (ya que las cintas y los discos son parte del *hardware*)

- Servicios

Tabla 10. Ejemplo de categorías referentes a servicios		
Ejemplos	**Problemática**	**Solución**
-Afinación de automóviles, servicio de -Grúas, servicio de	Este tipo de categorías contribuye a la ambigüedad. Muchos de estos servicios podrían ser parte de catálogos de las categorías. Sin embargo, en ocasiones la fuerza de la costumbre será la que rija el criterio para su creación.	De acuerdo al modelo de hiperonimia/hiponimia, la categoría que podría *adoptarlos* como hijo podría ser *talleres automotrices*, pero estos servicios son más conocidos y buscados de esta manera
-Servicios contables y administrativos		Su categoría padre sería *despachos contables y administrativos*; tendría que asegurarse que se conocen de esta manera

3.3.2.2 Propuesta de categorías prototípicas

El criterio para la creación de las categorías consistió en mantener un *punto medio* entre lo general y lo específico. Para lograr lo anterior, se realizó una jerarquía de relaciones léxico-semánticas basada en la hiponimia e hiperonimia. Se tomó como modelo el concepto de *Basic Level Term* fundamentado en la representación mental que el ser humano tiene de cierta entidad, por ejemplo para una exclamación como *¡qué feo perro!* es más aceptada que *¡qué feo animal!* (demasiado general) o *¡qué feo xoloescuincle!* (demasiado específica). Rosch comprobó en su experimento que los términos básicos poseen una serie de rasgos recurrentes, como ser palabras cortas (*rojo*, en relación al *color*), muy comunes (*carne*, en lugar de *aguayón*) y utilizadas en contextos neutros. De alguna manera, los términos básicos son aquellos que vienen a la mente de los hablantes enseguida.

Para el proyecto del directorio de RedCúbica S.A. de C.V., se buscó un punto medio entre lo general y lo específico; en casos particulares se decidió la creación de una categoría madre a la que le fueran asociadas categorías hijas sí y sólo sí tuvieran una relación intrínseca, además se establecieron criterios editoriales para resolver ambigüedades respecto a la forma de nombrarlas (uso de palabras conflictivas como *producto* y *artículo*, por ejemplo). De igual manera se buscó que todas las categorías tuvieran importancia comercial. Este factor extralingüístico influyó en la decisión de crear más o menos especificidad, así como buscar las diferencias formales entre unas y otras. Muchas veces se cuestionó la verdadera esencia de las categorías, esto es, si aquello que se nombraba existía en el mundo real. La siguiente tabla muestra ejemplos de cómo se decidió renombrarlas para su clasificación en el directorio de RedCúbica S.A de C.V.; conviene aclarar que la información señalada con asteriscos es la aportación lingüística y el factor que distingue al directorio de la empresa de los otros:

Tabla 11. Propuesta de categorías prototípicas 1

Sección amarilla	***Páginas útiles*, *Yahoo!***	***Kompass***	**Directorio RedCúbica S.A de C.V**
Forrajes, depósitos	Sin categoría relacionada	Pienso y Forraje	-Alimentos para animales de granja -Alimentos para ganado *Aprovechamiento de palabras clave, como *pienso*, *forraje*, *granja*

-Bordados computarizados -Bordados electrónicos -Bordados mecánicos -Talleres de encajes de bolillo y bordados	Bordados	Bordado y encaje	-Servicios textiles -Servicios de bordado *Creación de categorías que siguen el modelo de *Basis Level Term*
Zapatos de seguridad industrial - Fábricas	Zapatería	Calzado de trabajo, de protección y para usos especiales	-Zapatos para uso industrial -Zapatos para trabajar *Categorías que siguen un modelo de *Basic Level Term*, lo que evitaría la redundancia o el rebuscamiento
-Frutas -Frutas secas -Jugos de frutas	Frutas	-Fruta seca elaborada -Frutas y verduras enlatadas, en frasco o en otros envases	-Frutas -Frutas secas -Bebidas de jugos naturales *Categorías que son parte de un nicho de mercado, pero que no llegan a un grado de especificación como el que muestra el directorio *Kompass*

Por razones de espacio, a continuación sólo se presentará una tabla comparativa de veinte categorías creadas para el directorio RedCúbica S.A de C.V

Tabla 12. Propuesta de categorías prototípicas 2

Sección amarilla	**Páginas Útiles, Yahoo!**	**Kompass**	**Directorio RedCúbica**
-Academias de personalidad véase *escuelas e*	-Escuelas de modelaje, véase también *institutos y*	-Escuelas de modelos -Agencias de modelos	-Agencias de modelos -Escuelas de modelaje *Algunas veces la

institutos de modelos y personalidad -Agencias de modelos	*universidades* -Agencias de modelaje		costumbre, a partir de la mayor frecuencia de uso de una voz, determinará los nombres de las categorías
-Abanicos – transportadores véase *transportadores*	Sin categorías relacionadas	-Transportadores amortiguados -Cadenas transportadoras -Cintas transportadoras -Transportadores lineales multieje	-Artículos para uso industrial -Insumos para artículos de uso industrial -Artículos para transportación de materiales *Se prefiere un criterio que vaya de lo general a lo particular
-Agencias de publicaciones *Los anuncios que aparecen dentro de esta categoría refieren a empresas dedicadas a la publicación de textos	-Publicaciones, agencias *No se hizo un análisis sobre las categorías y tienen en el mismo error que el directorio de la *Sección amarilla*	-Publicaciones *on line* -Editores de publicaciones de agricultura -Editores de publicaciones por entregas -Publicaciones económicas y de negocios -Editores de publicaciones profesionales sobre informática	Editoriales *Esta categoría tendría como palabras asociadas a *publicaciones* y las que deriven de su giro comercial (*libros, librerías, editores, editor, editar, entre otras*), pero nunca *agencia*

		-Periódicos y revistas *Demasiada especificación deriva en confusión	
-Aditivos para concretos -Materiales ligeros para construcción -Materiales para construcción	-Concretos, materiales y servicios -Materiales para construcción	-Aditivos hidrófugos e hidrorepelentes para asfalto	-Materiales para construcción -Ferreterías -Tlapalerías
-Alquiler de sillas -Fiestas en general -Organización de banquetes a domicilio y salones para	Sin categorías relacionadas	-Lonas, alquiler -Alquiler de equipos para fiestas, festivales y ocasiones especiales	-Alquiladoras -Organización de eventos sociales
-Automóviles, agencias y compra - venta	-Automóviles, compra y venta	-Concesionarias de vehículos -Automóviles usados (concesionarias)	-Agencias automotrices *Sinónimo: Concesionarias
-Bufetes jurídicos	-Abogados	-Abogados y procuradores para la industria de la construcción *Demasiada específicidad	-Despachos jurídicos *Sinónimos: -Despachos legales, -Despachos de abogados
-Editores -Editoriales	Sin categorías relacionadas	-Editores de catálogos -Editores de libros	-Editoriales
-Fotografía	-Fotografía comercial	-Fotografía para moda -Fotografía de arquitectura -Fotografía digital	-Fotógrafos
-Estaciones difusora	-Televisoras	-Emisoras de radio y	-Televisoras

de televisión		televisión	
-Helados véase nieve y helados – Fábricas paleterías	Sin categorías relacionadas	-Helados (comercio) -Helados y sorbetes -Instalaciones para heladerías	-Heladerías -Paleterías *Categorías sinónimas
-Margarinas y mantequillas	Sin categorías relacionadas	-Margarina -Margarina comestible -Margarina líquida	-Productos alimenticios -Productos derivados de la leche
-Pantalones - fábricas de	-Ropa véase también *ropa* *ropa interior para mujer* *ropa para damas* *ropa para niños*	-Pantalones cortos para mujer -Pantalones, pantaloncillos, jeans (comercio)	-Ropa - fábricas de -Ropa interior – fábricas de
-Regalos - envolturas, bases y accesorios	Sin categorías relacionadas	-Artículos de regalo (comercio) -Regalos de cerámica -Papel para envolver regalos	-Mercería y regalos categoría sinónima: - Tienda de regalos palabra clave: regalos
Reparación de calzado véase talleres de reparación de calzado	Sin categorías relacionadas	Reparación de calzado y duplicación de llaves, en grandes almacenes o en tiendas detallistas	Reparadoras de calzado categoría sinónima: -Talleres de reparación de calzado
-Taquerías y torterías	Sin categorías relacionadas	Sin categorías relacionadas	-Torterías -Taquerías -Pulquerías
-Tortillas, expendios de	Sin categorías relacionadas	-Panadería fresco categoría similar	-Tortillerías categoría sinónima:

			-Tortillas, expendios de
-Talleres mecánicos -Talleres electromecánicos	-Talleres mecánicos	-Vehículos	-Talleres automotrices categoría sinónima: -Talleres mecánicos
-SPA Salud por Agua véase también *clínicas de belleza* -Estética femenina clínica de -Estética masculina – clínica de	-SPA -Salones de belleza	-Centros de salud Ayurveda y SPA -Salones de belleza -Peluquerías para señoras -Peluquerías para caballeros -Salones de belleza y peluquerías	-SPA's *Categoría sinónima: -Centros de belleza -Salones de belleza Categoría sinónima: Estéticas **Peluquerías **Sólo se podrán clasificar en esta categoría cuando den servicio exclusivo a hombres

3.3.3 Propuesta lingüística integral

Se formuló el desarrollo de herramientas lingüísticas que aportaran un valor agregado al algoritmo de búsqueda y recuperación de información. Como se ha mencionado, se proyectó un tiempo para analizar e identificar problemáticas de los directorios comerciales en línea; posteriormente se hicieron las propuestas pertinentes para la creación de categorías y recursos lingüísticos. Se estimó que el desarrollo de éstas fuera de un año a partir de abril del 2006; sin embargo, debido a la magnitud del proyecto y a la rotación del personal, los objetivos alcanzados al cierre de la empresa se limitaron al análisis e identificación de problemas y a la creación de categorías. El desarrollo de herramientas lingüísticas sólo se esbozó, para lo cual se llevó a cabo una investigación sobre cómo desarrollar lematizadores, analizadores morfológicos, desambiguadores semánticos, además de la inclusión de los recursos léxicos pertinentes para una integración lingüística de este tipo. En las siguientes páginas solamente se describirán las propuestas y la metodología más conveniente para su desarrollo.

3.3.3.1 Herramientas lingüísticas: lematizador, analizador morfológico, analizador semántico (desambiguador)

3.3.3.1.1 Lematizador

La propuesta lingüística presentada incluyó el desarrollo de un lematizador; esta herramienta tiene como objetivo encontrar la raíz de las palabras para poder encontrar más palabras a partir de ella. Como ya se ha explicado en el apartado 2.2 referente a las herramientas de la lengua usadas en ingeniería lingüística, lo anterior amplía los resultados encontrados ya que el algoritmo de búsqueda no se limita a encontrar sólo la palabra escrita en la caja de búsqueda, sino también a sus diferentes formas flexionadas. Los pasos para su desarrollo son los siguientes:

- Identificación del tipo de sufijos que se apegan a las reglas léxicas del español (prefijos y sufijos, marcas de género [femenino y masculino] y de número [plural y singular])
- Identificación de patrones y excepciones
- Análisis de los algoritmos existentes respecto al tema

Las reglas léxicas aplicables son las siguientes:

- Ubicar los prefijos y sufijos que ocurren frecuentemente en español (muchos de los algoritmos sólo quitan sufijos por la uniformidad que representa este procedimiento)

Tabla 13. Ejemplo del lematizador

Prefijos	**Sufijos**	**Marcas de género**	**Márcas de número**
a	ura	Gat**o/a**	Lápi**z**/lápic**es**
ab	miento	Muñec**o/**a	Rat**ón**/raton**es**
abs	anza		
anti	ción		
per	sión		

- Identificar los sufijos que ocurren juntos (*cion* +*es*, especificaci**ones**)
- Establecer el orden en que ocurren (sufijo +número)

Las reglas para identificar patrones y excepciones son:

- Dos sufijos que ocurren juntos no pueden pertenecer a la misma clase

- Las reglas que quiten sufijos de más al final de cada palabra deben ser procesados en un paso anterior a los que quitan otros
- Si un sufijo aparece siempre que ocurra otro, este sufijo es condicional a la aparición del anterior
- Las palabras terminadas en *r* suelen quedar con distinta raíz. Por ejemplo, *caminar* y *caminando*, por lo cual primero se debe eliminar – el sufijo *ndo* y posteriormente la *r*
- Siguiendo el ejemplo anterior, las palabras que terminan con vocales se dejan para el final cuando su raíz termine en consonante, por ejemplo, las palabras *terminación* y *terminal* y *terminó.*
- En último paso, se aplica una tercera regla que elimina las tildes de la raíz resultante. Por ejemplo, en *diálogo* y *dialogó*.

Hay además reglas propias del español. Por ejemplo, el sufijo *-nos*, que no es sufijo en palabras como *campesinos*, *casinos*, *caminos*, etc.; pero sí en *hacernos*, *ponernos*, *presentarnos*, etc.

Se pensó que el algoritmo más convencional para este tipo de aplicaciones era el de Porter; éste se desarrolló en 1979 para el idioma inglés, aunque se ha adaptó para el idioma español. Este algoritmo permite hacer *stemming* [48] , esto es, extraer los sufijos y prefijos comunes de palabras diferentes pero con una misma raíz. Básicamente, el lematizador genera las formas de inflexión de una palabra determinada en función de las reglas (por ejemplo, *corriendo*, *corrió*, *corredor*, son formas de la palabra *correr*). Como se ha mencionado, el uso de esta herramienta lingüística permite más resultados, además de proporcionar mayor precisión al momento de búsqueda.

3.3.3.1.2 Analizador morfológico

El desarrollo de un analizador morfológico tiene dos objetivos: por una parte sirve al lematizador para construir las palabras, y por otra parte analiza morfosintácticamente cada elemento dentro de las oraciones. Los pasos para desarrollarlo son:

- Identificación de morfemas (previamente se debió integrar una base de datos léxicos que los incluyan)
- Aplicación del lematizador y las reglas para construir/deconstruir palabras
- Interpretar a qué categoría gramatical corresponde cada morfema identificado (previamente se debió anotar un corpus léxico)

[48] *Stemming*: método para reducir una palabra a su raíz, un *stem* o tema. El *stemming* es especialmente útil para sistemas de recuperación de información. Su uso aumenta el número de documentos que se pueden encontrar en una consulta (Barité, 2011).

3.3.3.1.3 Analizador semántico (desambiguador)

Esta herramienta lingüística fue la única de la que se desarrollaron los primeros pasos. Se inició con la anotación del corpus de palabras del índice de clasificaciones hecho para RedCúbica S.A. de C.V. Cada una de estas palabras se debía calificar como *Normal*, *Parcial*, *Ambigua* o *Neutra*. Se integró además una lista de sinónimos. A partir de lo anterior, el algoritmo ampliaba o restringía el rango de resultados obtenidos, por ejemplo entre dos palabras discernía por el contexto de búsqueda (*ratoneras* como *trampa para ratones* y *ratoneras* como *jaula para ratones*). La siguiente tabla muestra los resultados que deberían arrojarse dependiendo de las combinaciones obtenidas:

Tabla 14. Ejemplo del analizador semántico (desambiguador)

Ruta de decisiones		**Solución**
Normal + normal	Zapatos +futbol	Las dos palabras que integran esta categoría tienen bien definidos sus contextos referenciales, por lo que no existe posibilidad de ambigüedad (los zapatos siempre serán zapatos e igualmente el futbol)
Parcial + normal	Artículos + oficina	La palabra *artículos* se considera parcial porque es demasiado general y puede dar lugar a la confusión. En estos casos sólo se restringe la aparición de resultados en los que aparezca la palabra *oficina* (la palabra artículos puede sustituirse por *productos*, *Insumos, equipo)*
Ambigua + normal	Plantas + ornato	La palabra *plantas* tiene varias acepciones (como vegetal, parte del cuerpo o establecimientos); es necesario definir los contextos en los que pueda aparecer esta palabra para restringir la aparición de

Ambigua + normal	Plantas + ornato	La palabra *plantas* tiene varias acepciones (como vegetal, parte del cuerpo o establecimientos); es necesario definir los contextos en los que pueda aparecer esta palabra para restringir la aparición de resultados

Este desambiguador se quedó a mitad del desarrollo; si bien ha sido superado en muchos sentidos fue práctico en su momento y se ajustaba a los requerimientos del sistema.

3.4 Recursos léxicos (redes léxico-semánticas [sinónimos, hipónimos, hiperónimos, palabras clave], léxicos computacionales y ontologías)

Para el desarrollo integral del sistema lingüístico se necesitaba la inclusión de recursos léxicos, como corpus computacionales, redes léxico-semánticas y ontologías. Estos recursos son parte fundamental para la estructuración del conocimiento léxico, morfosintáctico y semántico de la información y en consecuencia para la recuperación de información. A continuación se describirán los rasgos generales de cada uno de ellos y la aplicación concreta al sistema lingüístico de RedCúbica S.A de C.V.

3.4.1 Antecedentes teóricos

María Antonia Martí trata en su libro *Lexicografía computacional y semántica*[49] sobre el problema que implica la aplicación de los recursos léxicos en el Procesamiento del Lenguaje Natural. Explica que entre los recursos léxicos que sirven de apoyo al PLN se encuentran diccionarios electrónicos, léxicos computacionales, taxonomías y ontologías. Los objetivos de estos recursos son la codificación del significado de las palabras para satisfacer las necesidades propias del aprendizaje de la lengua, el procesamiento de la lengua y la experimentación psicolingüística. El objetivo de esta codificación será analizar en términos de metalenguaje la representación del conocimiento léxico. Los problemas centrales al respecto se centran en establecer criterios para el tratamiento de las unidades léxicas, así como su delimitación, caracterización y representación. Para lo anterior se emplean las redes léxico-semánticas y ontologías.

3.4.2 Redes léxico-semánticas

Se planteó el desarrollo de una base léxica a partir de las palabras que integraban el índice de clasificaciones hecho para el directorio comercial de RedCúbica S.A de C.V. Este *corpus* era manejable por su tamaño y permitiría la experimentación para un desarrollo posterior a mayor escala ya que al ser pequeño sería óptimo para su análisis morfológico, sintáctico y semántico. Las redes léxico-semánticas estructuran las palabras en función de la información semántica obtenida a través de las relaciones de sinonimia, antonimia, homonimia, hiperonimia, hiponimia o meronimia (conceptos explicados con anterioridad en la nota a pie número 38); por ejemplo, la palabra asociada a la categoría *zapaterías* sería *zapatería*, siendo ésta el hiperónimo de *zapaterías ortopédicas*, *zapaterías deportivas*, y merónimos como *zapatos*, *tenis*, *ballerinas*,

[49] Martí, M.A. *Lexicografía computacional y semántica*. Barcelona: Edicions Universitat de Barcelona, 2003.

sandalias, entre otros. Además se incluirían extranjerismos (*soccer shoes*) y palabras clave (*tiendas de zapatos, zaptos* [errores ortográficos]) directamente relacionadas con la categoría en cuestión. Esta clase de recursos servirían para la construcción de una ontología, además de aprovecharse como referencia para la anotación semántica del corpus. Se tomaría como referencia *WordNet*, proyecto desarrollado por la Universidad de Princeton. A partir de esta base léxica se implementaría un corrector ortográfico (no desarrollado aún por la *Sección amarilla*) que subsanara los errores ortográficos cometidos por los usuarios. De esta manera, el sistema lingüístico sería una herramienta integral para la recuperación de información en ambiente web.

3.4.3 Léxicos computacionales

Los léxicos computacionales incluyen información morfológica, sintáctica y semántica y no sólo la definición de la palabra (como usualmente aparece en la entrada de un diccionario). El uso de un léxico computacional en el directorio de RedCúbica S.A de C.V. sería una mejora para las búsquedas realizadas por los usuarios mediante expresiones en lenguaje natural (por ejemplo, escribir en la caja de búsqueda la pregunta ¿dónde encuentro un cerrajero en domingo?) y no sólo por medio de palabras clave (cerrajero, domingo, abierto). Continuando con el ejemplo anterior, en la categoría *zapaterías* se indicaría que se trata de un nombre, así como su género femenino y el número plural. De esta manera, se asociaría a la categoría la forma léxica *zapatería* (al respecto, consultar el apéndice 1 referente a las reglas generales para la creación de categorías; se encontrará la regla sobre el uso de plurales). Cabe señalar que también se asociaría a este lema su correspondencia en inglés (*shoe store, shoe shop*) y se restringiría a ciertos contextos su traducción automática para mejores resultados.

3.4.4 Ontologías

La parte esencial en todo sistema de recuperación y extracción de información es la representación del conocimiento. Los datos deben estar estructurados y organizados de tal manera que sea posible un acceso orientado a los contenidos. Esta estructuración del conocimiento es una especie de conceptualización y sirve para la extracción automática. Una ontología es una descripción formal de un conjunto de conceptos relacionados dentro de un dominio concreto. La ontología forma un cuerpo de conocimiento organizado en categorías definidas y relacionadas entre sí en función de sus correspondencias. Para el directorio de RedCúbica S.A. de C.V. era importante este proceso porque cada negocio o servicio tendría asociado un catálogo de productos ofrecidos. Así, la ontología se estructuraría de la siguiente manera:

- Conceptos: se trata de la descripción de las categorías que integran el dominio. Por ejemplo, en la categoría *vinatería* cada producto representaría una clase: *tequilas*, *brandys*, *vodkas*, *rones*, etcétera. A estas clases se le podrían asignar más niveles distintivos, llamados subclases. Esta estructuración de la categoría la convertiría en una taxonomía jerárquica. Como se observará, la correspondencia entre la representación del conocimiento y la elaboración de categorías es similar al establecer niveles hiperónimicos e hiponímicos.
- Ejemplares: se trata de entidades concretas de las clases o subclases, como marcas.
- Relaciones: es una representación de los diferentes tipos de conexión que se establecen entre clases y subclases. Responden al modelo meronímico (ser un tipo de, ser una parte de)
- Propiedades: son los atributos particulares de las clases y subclases. En el ejemplo de *vinatería* se establecerían las particularidades sobre el origen de los productos (si son de importación o exportación), el año de la cosecha (en el caso de los vinos), el precio.

Este tipo de estructuración de información sería de gran utilidad para el usuario, ya que agilizaría los procesos de búsqueda y selección de los resultados.

Conclusión

El florecimiento de las nuevas tecnologías ha supuesto la apertura de nuevos campos de acción en torno a la ingeniería lingüística y el procesamiento del lenguaje natural. Lo anterior ha llevado a los profesionales de la lengua a descubrir un nuevo panorama laboral prometedor en el que los más preparados serán aquellos que puedan adaptarse a equipos de trabajo multidisciplinarios.

El objetivo principal de este informe académico fue dar a conocer el trabajo hecho en la empresa RedCúbica S.A de C.V., como pionera en el desarrollo de software en México. Para una comprensión global del tema se decidió mostrar un panorama general del tema tecnológico y sus repercusiones en la sociedad de la información, además de exponer las áreas de investigación, que todavía necesitan desarrollarse en torno a las nuevas tecnologías.

Estas nuevas tecnologías, en años recientes, han evolucionado a grandes pasos: hoy se habla de una *Web Semántica*, de la *Web 3.0*, del campo de minería de datos y las posibilidades de negocios que ofrece la explotación de estos recursos (como *Fourth Square*), sin embargo en el año 2005 la inclusión de factores lingüísticos ya estaba entre los proyectos a desarrollar por la empresa RedCúbica S.A. de C.V. En este sentido, la empresa fue visionaria al adelantarse a su tiempo e incluir este tipo de herramientas muchos años antes que cualquier otra en México. El trabajo en la empresa es una demostración de cómo las ideas intuitivas sobre la lengua que tienen ingenieros y programadores no siempre corresponden a la teoría lingüística que un profesional en el área tiene. En este sentido, se debe enfatizar la flexibilidad que deben tener los lingüistas para adaptarse a las necesidades prácticas que implica el desarrollo de software. El objetivo del trabajo lingüístico será el de un guía que aporte ideas sobre cómo analizar de la manera más óptima el lenguaje sin perder de vista la parte formal y teórica del mismo.

Al respecto, si bien la metodología utilizada en RedCúbica S.A. de C.V.no tuvo un rigor formal propio de investigaciones lingüísticas, sí cumplió con el objetivo de desarrollar un sistema integral (aunque rudimentario) para recuperación y extracción de Información en ambiente web.

Siempre será necesario hacer hincapié en la necesidad de integrar a los planes de estudio las realidades que la sociedad demanda. La tecnología se ha convertido en parte esencial para la sociedad; por ello son necesarios profesionales que la entiendan desde todos los ángulos posibles y no sólo desde el área de su especialidad. Lo anterior no sucederá si no se integran a los planes de estudio de la licenciatura de

Lengua y Literaturas Hispánicas materias que aumenten el interés de los estudiantes en las nuevas tecnologías. Asignaturas como PLN, ingeniería lingüística e inteligencia artificial darían a los alumnos nuevas plataformas de desarrollo profesional. Desde mi experiencia profesional, la preparación recibida en la licenciatura fue teórica más que práctica, y si bien los conocimientos adquiridos fueron fundamentales para el desarrollo del *software* lingüístico, el verdadero conocimiento sobre este tema se dio a partir de la investigación de referencias bibliográficas al respecto.

Hoy en el mundo existen empresas de desarrollo de software que dan oportunidades laborales a lingüistas interesados en proyectos de PLN e ingeniería lingüística. En México existen proyectos tanto gubernamentales como privados dedicados al tema. En este momento, gracias a la experiencia obtenida en RedCúbica S.A. de C.V., participo en un proyecto de análisis léxico-semántico enfocado a aplicaciones web. Espero que con este trabajo se evidencie las oportunidades que un lingüista puede tener en un país como México.

Apéndice 1

1. Reglas generales para la creación de categorías:

Respecto a la descripción de las categorías:

Regla 1: las descripciones siempre se escribirán claras, esto es, sin caer en tecnicismos que puedan crear más confusión.

Regla 2: si se usan de tecnicismos se deberán explicar y ejemplificar.

Regla 3: las palabras utilizadas en las descripciones de las categorías deberán ser entendibles sin ninguna explicación aparte.

Regla 4: siempre que una categoría resulte ambigua frente a otra se deberá hacer una referencia circular entre ellas para hacer notar sus diferencias.

Respecto a las preposiciones utilizadas en los rótulos de las categorías:

Regla 1: siempre se tenderá a omitir el uso de preposiciones en los rótulos de las categorías. Por ejemplo: *artículos de fotografía* / *artículos fotográficos*; *tiendas de frutas/ fruterías.*

Regla 2: si no se puede omitir el uso de preposiciones se preferirá siempre la preposición *de*: *tiendas de ropa*.

Regla 3: si el uso de la preposición *de* resulta ambiguo se preferirá el uso de *para: artículos para fiestas*.

Respecto a la los rótulos de las categorías:

Regla 1: los rótulos de las categorías siempre deberán ser entendibles, esto es, que al leerlos inmediatamente sepamos lo que se encuentra clasificado en ella.

Regla 2: los rótulos de las categorías siempre deberán hacer referencia a un tipo específico de establecimiento. Éstos serán: fábricas, talleres, tiendas, puestos, hospitales, clínicas, o profesionistas independientes.

Regla 3: cada uno de estos rótulos tendrá sinónimos.

Regla 4: los rótulos de las categorías siempre serán los más populares siempre y cuando no causen ambigüedad.

Regla 5: al hacer categorías se deberán omitir aquellas palabras que resulten redundantes: *empresas aseguradoras*/*aseguradoras*.

Regla 6: si resulta ambiguo el nombre de la categoría, se preferirá referir al tipo de establecimiento: *empresas de consultoría inmobiliaria*.

Regla 7: nunca se deberán hacer dos categorías que se refieran a lo mismo como *carpinteros* / *carpinterías*: siempre se elegirá el rótulo que pueda causar menos ambigüedad.

Respecto al uso de las palabras "*servicios*" y "*especializada*" en los rótulos de las categorías

Regla 1: la palabra "*servicios*" solo se utilizará acompañada de un objeto material (*servicios audiovisuales*) y no de una disciplina inmaterial (*servicios de publicidad*); en este último caso es redundante pues publicidad ya implica un servicio mientras que audiovisual no.

Regla 2: la palabra "*especializada*" no se usará en los rótulos de las categorías pues se entiende que "*tienda especializada en tornillos*" es lo mismo que "*tienda de tornillos*". La palabra se sobreentiende si se especifica un objeto en el rótulo de las categorías.

Respecto a la especificación de las categorías

Regla 1: la tendencia será la creación de categorías específicas siempre y cuando no resulten incoherentes o extremadamente específicas. Asimismo se crearán categorías particulares que puedan contener a estas categorías específicas; por ejemplo:

Nivel universal	Escuelas
Nivel general	Escuelas deportivas
Nivel particular	**Escuelas de artes marciales**
Nivel específico	**Escuelas de Karate**

Regla 2: en el caso de las profesiones y oficios la tendencia será la de crear categorías que abarquen no solo al profesionista que trabaje individualmente sino al que se asocie con otros y creen despachos, bufetes, etc. Por ejemplo:

Contadores	Despachos contables
Abogados	Bufetes jurídicos

Respecto al campo semántico de Salud:

Regla 1: se utilizará la palabra *hospital* para referirse a todos aquellos establecimientos que den asistencia médica y que proporcionen el servicio de cirugías mayores y estancia. Un hospital siempre será general o de algo en específico: *hospitales generales*; *hospitales pediátricos*.

Regla 2: el sinónimo de *hospital* será *sanatorio*.

Regla 3: se deberán distinguir como categorías a las *maternidades* y las *clínicas*; las primeras son hospitales acondicionados para asistir a las mujeres que dan a luz y las segundas son establecimientos que se dedican a atender especialidades y a realizar pequeñas cirugías, pero no tienen la infraestructura de un *hospital*.

Regla 4: para referirse a los *consultorios médicos* se preferirá usar el nombre común del médico especialista o general, puesto que suponemos que todos los médicos tienen un consultorio médico: *cardiólogo*.

Regla 5: se distinguirá como categoría a los *dispensarios médicos*.

Regla 6: no se omitirán las palabras *hospital*, *clínica*, *maternidades* ni *dispensarios médicos*.

Respecto al campo semántico de Producción:

Regla 1: siempre se llamará *fábricas* a aquellos establecimientos con las instalaciones y la maquinaria necesaria para producir, confeccionar u obtener productos en serie.

Regla 2: "*plantas*…" solo será el equivalente a la palabra *fábricas* en aquellos casos en los que así sean conocidas popularmente.

Regla 3: se usará la palabra "*taller*" solo para aquellos establecimientos en los que se elaboren artículos de manera artesanal. "*talleres de orfebrería*".

Regla 4: para referirse a los lugares en los que reparan artículos se preferirá la palabra "*reparadora*" siguiendo como ejemplo a las "*reparadoras de zapatos*". Por ejemplo: reparadora de electrodomésticos. Esta regla tiene como excepción "*talleres mecánicos*".

Regla 5: se tratará de omitir las palabras "*fábrica*", "*taller*" o "*plantas*" de los rótulos de las categorías siempre y cuando esto no cause ambigüedad. Por ejemplo: fábricas de *cervezas /cerveceras*; plantas armadoras de automóviles / armadoras de automóviles.

Respecto al campo semántico de comercio

Regla 1: a todo aquello que se dedique a la venta al consumidor final se le llamará *tienda*.

Regla 2: se hará una distinción entre *tienda* y *expendio* cuando este último solo se dedique a la venta de un determinado producto y su uso sea popular. Por ejemplo: *expendio de dulces*.

Regla 3: solo se usará la palabra *depósito* en los siguientes casos conocidos: *depósitos de cervezas* y *depósitos de vinos*.

Regla 4: se distinguirá a las *comercializadoras* como aquellas empresas dedicadas a vender la producción de las fábricas a las tiendas.

Respecto al campo semántico de servicios:

Regla 1: todo aquello que se dedique a la administración de servicios será llamado *empresa*.

Regla 2: lo anterior aplica todos sus correspondientes sinónimos.

Regla 3: una *empresa* siempre será de algo.

Regla 4: si es posible se omitirá la palabra *empresa* siempre y cuando no cause confusión.

Regla 5: solo se usará la palabra *agencia* como sinónimo de *empresa* cuando así sea conocida popularmente.

Respecto al campo semántico de profesionistas:

Regla 1: se llamará por el nombre popularmente conocido a todas aquellas personas que ejerzan su profesión de manera independiente. Los oficios entran en esta categoría.

Respecto al campo semántico de mercancías:

Regla 1: la palabra *artículo* solo se usará para designar a objetos concretos.

Regla 2: La palabra *producto* solo se usará para designar a objetos abstractos o que no sean contables.

Regla 3: se deberá a especificar la clase a la que un producto pertenece (*accesorio*, *insumos*, *equipo*, *materia* o *material*). Estos sustantivos son bien identificados, por lo cual sobran las definiciones.

Respecto al campo semántico de educación:

Regla 1: siempre se llamará *escuela* a todo establecimiento que enseñe algo. En este sentido no se hará diferencia entre lo público y lo privado, lo oficial, lo técnico o lo artístico.

Regla 2: cuando una *escuela* sea más conocida de otra manera se preferirá ésta (como *Universidad*).

Regla 3: sus sinónimos serán *academia*, *colegio* e *instituto* en las acepciones correspondientes.

Respecto al campo semántico de entretenimiento:

Regla 1: se hará una distinción entre *bares*, *cantinas* y *cervecerías*. La palabra *taberna* se usará como palabra clave de las tres categorías.

Regla 2: antros será el nombre de la categoría; se usará como palabra clave discotecas (palabra considerada pasada de moda)

Regla 3: serán considerados "*centros nocturnos*" aquellos lugares que ofrezcan algún tipo de variedad o espectáculo de entretenimiento que pueda ser considerado familiar.

Regla 4: serán considerados "*salones de baile*" a aquellos establecimientos que presenten música en vivo (géneros latinoamericanos como boleros, salsa, cumbia, etc.) y que tengan una pista de baile.

Reglas arbitrarias

Regla 1: la palabra "*asociaciones*" se usará para designar a los grupos de personas que busquen fines y objetivos comunes y pueden ser en torno a una profesión o una actividad.

Regla 2: la centro" solo se utilizará para referirse a aquellos lugares que reúnen a grupos de personas.

Regla 3: la palabra "*instituto*" solo se utilizará para referirse a aquellos lugares en los cuales se lleven a cabo investigaciones científicas, tecnológicas, médicas, culturales, sociales o artísticas.

Regla 4: Se preferirá no usar guiones en las palabras que conformen los rótulos de las categorías.

Apéndice 2

Documentación de la metodología del sistema

De acuerdo con este análisis, el directorio de Ubícame ha desarrollado una tipología de clasificación que organiza los datos de acuerdo a su naturaleza y características:

1. Sustantivos.
2. Giros.
3. Categorías.
4. Categorías determinantes.
5. Categorías por denominación.
6. Clasificaciones

Y como herramientas que pueden facilitar y complementar la búsqueda:

1. Sinónimos
2. Palabras clave
3. Sugerencia de búsqueda

Esta tipología de elementos funciona gracias al factor semántico que hace posible las relaciones semánticas y en consecuencia la herencia de palabras; de esta manera el directorio garantiza encontrar lo que sea como sea que se busque pero siempre con resultados de calidad.

Vamos a explicar los elementos que componen el sistema, la manera de extraerlos y las reglas para formarlos.

Sustantivos

El sustantivo es el elemento básico del sistema pues denomina a los objetos (productos, servicios, lugares y comercios).

Clases de sustantivos

Existen dos clases de sustantivos:

Sustantivos		
Clases de sustantivos	Ejemplos	Explicación
Simples	Alfombras	Un sustantivo único
Conceptuales	Pisos de adoquín	Un sustantivo conceptual es el resultado de la unión de dos o más sustantivos simples.

Obtención de sustantivos

Un sustantivo sirve para crear categorías

El proceso para la obtención de sustantivos es el siguiente:

- Selección de las fuentes; las más útiles son los índices de clasificaciones de los directorios (pues se refieren particularmente a negocios, profesionistas, lugares y puntos de referencia).
- Separación de los sustantivos.
    - Criterios para separar sustantivos
        - ❖ Deben ser importantes por sí mismos: un sustantivo deberá tener una realidad concreta, coherente y aplicable para el directorio (relación con los negocios, lugares, puntos de referencia, personas y profesionistas):

Índice de clasificaciones		
Categorías	**Obtención de los sustantivos**	**Explicación del filtrado**
Fibra de vidrio	Fibra de vidrio	Se separa el sustantivo Vidrio porque es relevante como objeto/producto para el directorio
	Vidrio	
Fibras	Fibras	No se separaron los sustantivos “Aseo, Químicas, Sintéticas, Textiles” porque solos no son relevantes para el directorio.
Fibras para aseo	Fibras para aseo	
Fibras químicas	Fibras químicas	
Fibras sintéticas	Fibras sintéticas	
Fibras textiles	Fibras textiles	

        - ❖ No se separarán sustantivos simples demasiado específicos si existe un sustantivo más general que pueda abarcarlos.

Índice de clasificaciones		
Categorías	**Obtención de los sustantivos**	**Explicación del filtrado**
Conchas, caracoles y adornos marinos	Adornos (marinos)	Los sustantivos “Conchas” y “caracoles” ya están contemplados con el sustantivo “Adornos marinos” por lo tanto no es necesaria su separación (a memos que se verifique su importancia como categoría).
	Conchas✖	
	Caracoles✖	
	Adornos	El sustantivo “Adornos” es más general y se puede referir a otro tipo de adornos, no solo a los marinos, por eso se ha separado.

- Se separarán sustantivos compuestos que sean coherentes y lógicos.

Índice de clasificaciones		
Categorías	**Obtención de los sustantivos**	**Explicación del filtrado**
Alquiler de camas tipo hospital y sillones para inválidos	Camas tipo hospital Sillones para inválidos Camas Sillones Hospital	1. "Alguien" busca " venta de camas tipo hospital"; encuentra una categoría: "Camas tipo hospital y sillones para inválidos – Venta de" 2. Se comunica…. 3. Pregunta por las camas tipo hospital… 4. Le responden que ahí solo se venden "Sillones para inválidos". 5. Ese alguien perdió su tiempo… En el directorio de Ubícame esto no sucederá, porque quien venda solamente "Camas…" podrá clasificarse en esa categoría; quien venda "Sillones…" podrá clasificarse en esa categoría; quien venda los dos se deberá clasificar en ambas. Este proceso garantiza la precisión en los resultados.

- La separación redundante de sustantivos compuestos daña la rapidez y precisión del directorio.

Índice de clasificaciones		
Categorías	**Obtención de los sustantivos**	**Explicación del filtrado**
Conos- Barquillos para Nieve y Helados.	Barquillos para helados Barquillos para nieves Conos para helados Conos para nieves Barquillos Conos Helados Nieves	1. Barquillos y conos son sinónimos 2. Entonces "Barquillos para helados" y "Conos para helados" resultan ser lo mismo 3. La separación de estos dos sustantivos resulta redundante pues ambos se refieren a lo mismo. 4. En igual situación se encuentran los sustantivos "Helados" y "Nieves" 5. En ambos casos se debe decidir cuál palabra es la más apropiada para nombrar cada cosa.

Criterios de los sustantivos

1. Especificidad de los sustantivos: Los sustantivos del sistema serán generales; cuando se desee hacer una especificación del sustantivo se pondrá entre paréntesis:

Sustantivos
Cereales
Cereales (con fruta seca)
Cervezas
Cervezas(de trigo)

2. Definición de sustantivos: Cuando el sustantivo sea poco usual se deberá dar una definición; está se escribirá entre corchetes.

Sustantivos
Sorgo [Planta forrajera]
Alforjón [Trigo sarraceno]
Pantallas de cristales líquidos [LCD]

3. Palabras homónimas: Cuando existan sustantivos homónimos se hará una definición que pueda distinguirlos.

Palabras homónimas	Distinción
Clavos	herramientas
Clavos	Condimento alimenticio
Metales	Valores monetarios
Metales	Fierro, aluminio, acero
Discotecas	Salas para bailar
Discotecas	Venta de discos

La distinción de las palabras homónimas se pondrá al momento de crear categorías. Esta información se escribirá entre corchetes.

Ejemplo estándar para crear categorías			
Giro	Industria	Giro	Venta
Clasificación	Herramientas	Clasificación	Alimentos
Sustantivo	Clavos	Sustantivo	Clavos
Categoría		Categoría	
Sustantivo	Clavos	Sustantivo	Clavos
Giro	Fabricación	Giro	Venta
Categoría Determinante	0	Categoría Determinante	0

Por denominación	0	Por denominación	0
Categoría creada	Clavos [herramientas]-Fabricación de	Categoría creada	Clavos [Condimento]-Venta de

4. Extranjerismos

El español tiene preferencia frente a otros idiomas.

Extranjerismo	Traducción al español
hard disk	Discos duros
Aerobics	Aeróbicos
Block	Blocs
Jeans	Pantalones de mezclilla
Blazer	Chaquetas
Cassette	Casetes

En la lista de sustantivos se deberán poner los extranjerismos entre corchetes.

Sustantivos
Discos duros [hard disk]
Aeróbicos [Aerobics]
Blocs [Block)
Pantalones de mezclilla [Jeans]
Chaquetas [Blazer]
Casetes [Cassette]

Excepciones

En algunas ocasiones si el sustantivo es más conocido por la palabra extranjera se preferirá ésta.

Palabra en español	Extranjerismo
Perforaciones corporales	Body-piercing
Perros calientes	Hot dogs
Bufé	Buffet
Cabaré	Cabaret

En la lista de sustantivos se pondrá entre corchetes el sustantivo en español.

Sustantivos
Body-piercing [Perforaciones corporales]
Hot dogs [Perros calientes]
Buffet [Bufé]
Cabaret [Cabaré]

5. Número de los sustantivos

En todos los sustantivos se preferirá el plural por razones de uniformidad; sin embargo se deberá considerar que no todos lo aceptan.

Se pondrá el plural a los sustantivos que así lo permitan.

Sustantivos
Abanicos
Abarrotes
Abogados

Se dejará en singular los sustantivos que no puedan ir en plural bien porque no es su forma morfológica o porque la forma en plural cambia su significado.

Sustantivos
Azúcar
Aerocarga
Aeromodelismo
Ajedrez

Los sustantivos conceptuales también se escribirán en plural.

Sustantivos
Aparatos auditivos
Aceites lubricantes
Cocinas integrales

Se debe considerar que no todos los sustantivos simples de los sustantivos conceptuales pueden llevar el plural.

Sustantivos
Bandas de garantía
Ligas de aluminio
Cajas de cartón corrugado

Los extranjerismos que así lo requieran y acepten llevarán el plural en español.

Sustantivos
Casetes
Panties
Blocs

6. Adjetivos

Los adjetivos atribuyen cualidades a los sustantivos; esta característica no tiene relevancia para el sistema, puesto que su uso especifica demasiado a los sustantivos.

Sustantivos
Automóviles rojos
Acabados decorativos perfectos
Casas grandes
Cobranzas rápidas
Piedras preciosas (zafiro, topacio, rubí)
Plantas naturales
Plantas artificiales
Automóviles seminuevos
Automóviles nuevos

Se les quitarán los adjetivos y se encontrará una manera más general de nombrarlos.

Sustantivos	
Acabados decorativos	Joyería
Casas	Plantas
Cobranzas	Plantas de ornato
Chiles	Automóviles
Chiles	Joyería

7. Abreviaturas

Para evitar confusiones, nunca se usarán abreviaturas en el directorio.

Abreviaturas	
Abreviatura	Palabra
Prims.	Primarias
Prims.	Primas

Se deberá “deshacer” esa abreviatura y poner la palabra que correspondiente a la categoría correcta:

Una vez hecho esto, se deberá actualizar la lista de los elementos afectados por este cambio (sustantivos, giros, categorías, clasificaciones).

Giros

Los giros cumplen la función de determinar qué se hace con un sustantivo. A diferencia de los sustantivos, un giro sin la combinación de éstos pierde su significado práctico dentro del sistema.

Clases de giros

En el directorio de Ubícame se han distinguido tres clases de giros para fines de organización interna.

Clases de giros	Explicación	Ejemplo
Servicios industriales	Son giros que participan en los procesos de transformación de materias primas para hacer productos	Excavación
Servicios comerciales	Estos giros se caracterizan por ser las actividades económicas por medio de las que se obtienen productos.	Venta
Servicios profesionales	Estos giros se refieren a aquellas actividades proporcionadas por profesionistas.	Enseñanza
Giros derivados de los términos de las actividades comerciales	Algunos giros no serán propiamente verbos sino derivaciones de los términos que se utilizan en el medio comercial.	Mayoreo

Obtención de los giros

Los giros deberán pertenecer a la Industria, el Comercio y los Servicios. Se obtienen de tres formas:

1. Por medio de las categorías del índice de clasificaciones de los directorios:

Índice de clasificaciones		
Categorías	**Obtención de sustantivos**	**Obtención de giros**
Casas y terrenos- Compra y venta	Casas	Compra
	Terrenos	Venta
Calderas - Reparación	Calderas	Reparación
Barbacoa- Elaboración y venta	Barbacoa	Elaboración

2. Por medio de los sustantivos compuestos:

Índice de clasificaciones		
Categorías	**Obtención de sustantivos**	**Obtención de giros**
Escuelas de enseñanza artística	Escuelas de enseñanza artística	Enseñanza
	Escuelas	
Ligas de aluminio para fundición	Ligas de aluminio para fundición	Fundición
	Ligas de aluminio	
Equipos para laboratorios de la construcción	Equipos para laboratorios de la construcción	Construcción
	Laboratorios de la construcción	

3. Por medio de los lugares donde se realiza la acción:

Índice de clasificaciones		
Categorías	**Obtención de sustantivos**	**Obtención de giros**
Equipos neumáticos-Fábricas de	Equipos neumáticos	Fabricación de
Deshidratadoras	Deshidratadoras (máquinas)	Deshidratación de
	Deshidratadoras (plantas deshidratadoras)	

Criterios de los giros

Forma morfológica de los giros: por razones de uniformidad, todos los giros obtenidos tendrán la misma forma morfológica:

1. Se preferirá siempre la terminación en –ción

Presentación estándar de un giro
Fabricación

2. Cuando el giro no acepte esta forma, se presentará con terminación en vocal o en infinitivo:

Presentación alterna de un giro
Venta
Diseño
Compra
Alquiler

3. Para decidir la forma que llevarán los giros es necesario probar con todos los casos y decidir siempre por la que sea más correcta y conocida.

4. Giros ambiguos: Los giros, al igual que los sustantivos, pueden presentar ambigüedades que dificulten la precisión del sistema del directorio. Para desambiguarlos se seguirá el siguiente procedimiento:

Área Servicios comerciales

Giros en conflicto	Desambiguación	Decisión
Venta	Vender se refiere a la acción de exhibir mercancías o propiedades con fines monetarios.	Vender y Comercializar se refieren a la misma actividad, sin embargo éste último tiene mucha mayor complejidad pues indica a una venta a gran escala (venta de un producto en todo el país) mientras que Vender se refiere más bien a la actividad local; por lo tanto no son giros que puedan ser sinónimos.
Comercialización	Comercializar se refiere a la acción de hacer que un producto tenga una organización y unas condiciones comerciales para su venta	
Giros en conflicto	**Desambiguación**	**Decisión**
Renta	Rentar se refiere a producir cada cierto tiempo una cantidad de dinero o un beneficio	"Yo obtengo una renta por alquilar mi casa"; en este ejemplo los giros son correctamente utilizados; sin embargo en la vida práctica no existe tal distinción: "Yo rento sillas" es igual a "Yo alquilo sillas". Por lo tanto, estos giros se considerarán sinónimos. El siguiente paso es decidir cuál es la opción más correcta para aparecer en el directorio.
Alquiler	Alquilar se refiere a dar una cosa para sus uso por un determinado tiempo a cambio de una cantidad de dinero	

Esta desambiguación de giros similares se realizará para facilitar el trabajo al momento de crear las categorías.

5. Sustantivación de giros dentro de los sustantivos: Un giro dentro de un sustantivo compuesto adquirirá función y significado de sustantivo:

Índice de clasificaciones		
Categorías	**Obtención de sustantivos**	**Obtención de giros**
Escuelas de enseñanza artística	Escuelas de enseñanza artística	Enseñanza
	Escuelas	
Ligas de aluminio para fundición	Ligas de aluminio para fundición	Fundición
	Ligas de aluminio	
Equipos para laboratorios de la construcción	Equipos para laboratorios de la construcción	Construcción

Clasificaciones

Dentro del sistema debe existir un control que mantenga ordenados a los elementos que lo conforman. Este orden se logrará a través de clasificaciones**.**

Criterios para clasificar

En el sistema existe un único criterio para clasificar y es el de hiperonimia. Este criterio se ha elegido por ser la manera más sencilla de organizar semánticamente la información. La relación de hiperonimia es una red semántica en la cuál los conceptos están organizados en un subconjunto de elementos que forman parte de otro más general: a los conceptos superiores se le asigna uno o varios conceptos hijo, que a su vez tienen otros. En la relación de hiperonimia al concepto superior se le llama Hiperónimo y al concepto hijo Hipónimo

Hiperónimo	**Hipónimo**
Animales	Mamíferos
Mamíferos	Cánidos
Cánidos	Perros
Perros	Chihuahueños

Elementos que se clasifican en el sistema.

1. Giros

Clasificación de giros	
Servicios industriales	Servicios comerciales
Demolición	Exportación
Mantenimiento	Arrendamiento
Construcción	Venta
Refinación	Compra
Servicios profesionales	Giros derivados
Administración	Mayoreo
Adiestramiento	
Limpieza	
Vigilancia	

Los giros se clasifican de acuerdo a los tres criterios de los giros anteriormente explicados:

2. Sustantivos

Procedimientos para clasificarlos:

Se deben encontrar los sustantivos generales (Hiperónimos) que puedan agrupar a otros sustantivos (Hipónimos) de acuerdo con características comunes:

Clasificación de sustantivos
Jardinería
Herbicidas para invernaderos
Fertilizantes para jardines con pasto y flores

Plásticos para invernaderos
Pasto sintético para jardines
Plantas (naturales)
Plantas (ornamentales)
Flores (deshidratadas)
Macetas
Herbicidas
Orilladoras
Muebles para jardín
Fuentes ornamentales
Mangueras para riego
Podadoras para pasto
Flores artificiales
Invernaderos
Figuras para jardines
Equipo para jardinería
Muebles de forja
Equipo para jardín
Pasto artificial para jardines
Flores (naturales)
Viveros
Herbicidas para viveros
Plásticos para viveros

Dentro de esta clasificación se hará una subclasificación (Hiperónimos) de sustantivos (Hipónimos) que tengan compartan más características entre ellos:.

Lista de sustantivos
Jardinería
Mobiliario
Muebles para jardín
Muebles de forja
Fuentes ornamentales
Sombrillas para jardines
Sombrillas para parques
Figuras para jardines
Equipo para jardín
Mangueras para riego
Equipo para jardín
Equipo para jardinería
Podadoras para pasto

Orilladoras
Materiales para el mantenimiento de parques y jardines
Herbicidas
Pasto artificial para jardines
Fertilizantes para jardines con pasto y flores
Pasto sintético para jardines
Invernaderos
Invernaderos
Artículos para invernaderos
Herbicidas para invernaderos
Plásticos para invernaderos
Viveros
Viveros
Artículos para viveros
Herbicidas para viveros
Plásticos para viveros

Categorías

Las categorías son la presentación final de la información del sistema.

Clases de categorías

El sistema distingue tres clases de categorías:

Tipos de categorías	Descripción	Ejemplo
Categoría simple	La categoría simple se compone de sustantivo (s) y un giro.	Ganado – Cría de
Categoría determinante	La categoría determinante es aquel sustantivo que lleva implícito el giro.	Carnicería
Categoría “Por denominación”	Por denominación son aquellas categorías que no tienen giro explícito ni implícito aunque si indican una actividad determinada.	Profesores Médicos Contadores

Criterios de las categorías:

Las categorías adquieren su clasificación por medio del sustantivo que llevan (no importa si éste es simple o conceptual). Por ejemplo, la categoría por denominación “Arquitectos” se clasifica en “Profesionistas de la construcción”:

Arquitectos	Lista de sustantivos
Decoradores	**Profesionistas de la construcción**

Contratistas
Ingenieros civiles
Ingenieros contratistas
Contratistas
Ingenieros constructores
Ingenieros estructurales
Albañiles

Categoría
Cereales – Venta de {a granel}

Especificación de la categoría: la categoría podrá ser específica cuando indique una particularidad o modalidad que la caracterice. Esta caracterización deberá ir entre llaves {}:

Desambiguación de Categorías: Se escribirá una distinción para distinguir categorías similares pero que se refieran a situaciones distintas; se hará al momento de crear las categoría y se escribirá entre corchetes:

- En el caso de una categoría simple la distinción se tomará de los giros y los sustantivos que fueron seleccionados por ser demasiado parecidos; esta distinción es útil para que no haya equivocaciones al categorizar:

Categoría simple
Aguacate (hass) – Venta de [Tener u ofrecer algún bien o mercancía a disposición del público]
Aguacate (hass) - Comercialización de [Se refiere a la condición y vías de distribución que un producto tiene para su venta]

- Las categorías determinantes que se formen con los mismos sustantivos y giros se deberán distinguir para que no exista un error al categorizar:

Categoría determinante	
Sustantivo	Automóviles
Giro	Venta
Categoría Determinante	Agencias automotrices
Por denominación	No aplica
Categoría creada	Agencias automotrices [Esta categoría se refiere a las empresas que venden Automóviles nuevos].

Categoría determinante	
Sustantivo	Automóviles

Giro	Venta
Categoría Determinante	Lotes de Automóviles
Por denominación	No aplica
Categoría creada	Lotes de Automóviles [Esta categoría se refiere a aquellos lugares en los cuáles se venden Automóviles seminuevos].

Creación de categorías

Procedimientos

Se escoge un sustantivo y un giro ya clasificados.

Clasificación	Servicios Industriales
Giro	Fabricación
Clasificación	Jardinería
Sustantivo	Muebles para jardín

Se verifica que la combinación exista en el mundo real (esto se puede hacer a través de Internet o en una práctica de campo)

Categoría simple	
Sustantivo	Muebles para jardín
Giro	Fabricación de
Categoría Determinante	No aplica
Por denominación	No aplica
Categoría creada	Muebles para jardín – Fabricación de

Categoría determinante	
Sustantivo	Tacos
Giro	Venta
Categoría Determinante	Taquería
Por denominación	No aplica

Categoría creada	Taquería

Por denominación	
Sustantivo	Médicos
Giro	No aplica
Categoría Determinante	No aplica
Por denominación	Médicos
Categoría creada	Médicos

Índice de palabras

La lista de palabras contiene cada una de las unidades aislables de la cadena escrita. Se originó al separar los elementos del sistema (Sustantivos, giros, clasificaciones, categorías, palabras clave). Así:

Origen de la lista de palabras	
Sustantivo	Separación
Ligas de fundición para aluminio	Ligas
	De
	Fundición
	Para
	Aluminio

Giro	Separación
Alquiler de	Alquiler
	de
Construcción de	Construcción
	de
Mantenimiento de	Mantenimiento
	de

Clasificación	Separación
Recreación y tiempo libre	Recreación
	Tiempo
	Y
	Libre
Maquinaria para la construcción	Maquinaria
	Para

	La
	Construcción

Esta lista de palabras sirve para mantener un control de calidad respecto a la ortografía de los elementos del sistema. Se divide en tres clases:

Lista de palabras		
Clase de lista de palabras	Descripción	Función
Lista de palabras en uso	Son palabras que están en funcionamiento dentro del directorio de Ubícame.	Sirve para saber exactamente que porcentaje de las palabras creadas para el directorio están en uso.
Lista de palabras en desuso	Son palabras que existen en el directorio pero que no se han usado todavía.	
Lista de palabras desechadas	Son palabras incorrectas o que el sistema no reconoce.	Cuando en una búsqueda un usuario escribe “Kidditos” el sistema generará un reporte para que los analistas de contenido decidan si la palabra debe ir a la lista de palabras desechadas.

Herramientas de búsqueda

Para lograr los objetivos del sistema son necesarias herramientas auxiliares en el proceso de búsqueda. Éstas son:

Sugerencia de búsqueda (¿Usted quiso decir…?)

Sirve como un corrector ortográfico que el sistema proporciona cuando las palabras escritas en el cuadro de búsqueda se escribieron incorrectamente. La sugerencia de búsqueda le muestra al usuario la manera correcta de escritura. La corrección ortográfica se basa en un sistema de aproximación matemática entre palabras. Así:

- Un usuario escribe en el cuadro de búsqueda:

Hosptal	BUSCAR

- El sistema le sugerirá otra variante ortográfica:

0 resultados Quizás usted quiso decir ***Hospital***

- La palabra incorrecta genera un reporte; los analistas de contenidos mandan esta palabra a la lista de palabras desechadas para que se integre al sistema del directorio de Ubícame y la reconozca como incorrecta y no vuelva a hacer un reporte.

Lista de palabras desechadas
Hosptal
Comptadora

Sinónimos

En el sistema los sinónimos son una herramienta más por medio de la cuál se encuentran los elementos. Los sinónimos funcionan como un sistema para sustituir palabras y complementarlas. La lista de sinónimos tiene tres columnas:

- La primera tiene el término que se ha escogido por ser el más correcto de un grupo de sustantivos
- La segunda tiene los sinónimos totales de ese término correcto.
- La tercera tiene las palabras que pueden ser hipónimos o acepciones del término correcto.

Lista de sinónimos		
Término correcto	Sinónimos	Dudas
Abanicos	Abanico Abano Abanillo Ventalle	Aventador Paipay Flabelo Baleo Soplillo
Abarrotes	Abarrotes Comestibles Despensas	Colmados Ultramarinos Abacería Fardo
Abejas	Abejas Insecto	Reina Abejarrón Abejón Abejorro Zángano
Abogados	Abogados Licenciados	Leguleyo Defensor Asesor Protector Consejero Jurista Criminalista
Fertilizantes	Fertilizantes Abonos Abonaduras	Mantillo Humus Guano Estiércol Excremento Nitrato Boñiga Bosta

Clases de sinónimos: Hay dos clases de sinónimos

Sinónimos			
Clases de sinónimos	Descripción	Ejemplo	
		Término correcto	Sinónimos
Sinónimos totales	Son aquellos que pueden sustituir a la palabra correcta sin causar problemas de significado	Escuelas	Escuelas
			Academias
			Institutos
			Colegios
Sinónimos por uso	Son aquellas palabras que sin ser sinónimos se utilizan como si lo fueran	Discapacitados	Discapacitados
			Inválidos
			Minusválidos

			Personas con capacidades diferentes

Criterios de los sinónimos

- Precisión de los sinónimos: El sistema solo aceptará los sinónimos totales y los contextuales; los hipónimos o las acepciones solo son referencias léxicas que no es necesario documentar.
- Sinónimos de palabras homónimas: La lista de sinónimos no distingue palabras homónimas:

Lista de sinónimos		
Término correcto	Sinónimos totales	Dudas
Clavo Clavo	Tachuelas Alcayatas Escarpias Punta Clavillo	

- Es así que cuando se busque “Clavos” los resultados ofrecidos puedan referirse a clavos de olor o clavos para madera. Por esta razón, las palabras homónimas se descartarán de la lista de sinónimos. Los sinónimos de estas palabras se pondrán como palabras clave a las categorías correspondientes.

Elementos	Datos	Palabras clave
Clasificación	Industria	*[“”],[“”]*
Giro	Fabricación	*[“”],[“”]*
Clasificación	Herramientas	*[“”],[“”]*
Sustantivo	Clavos	*[“Tachuelas”],* *[“ Alcayatas], [“Escarpia”]*
Categoría		
Sustantivo	Clavos	
Giro	Fabricación	
Categoría Determinante	0	
Por denominación	0	
Categoría creada	**Clavos (Herramientas) – Fabricación de**	*[“”]*

Procedimientos de búsqueda

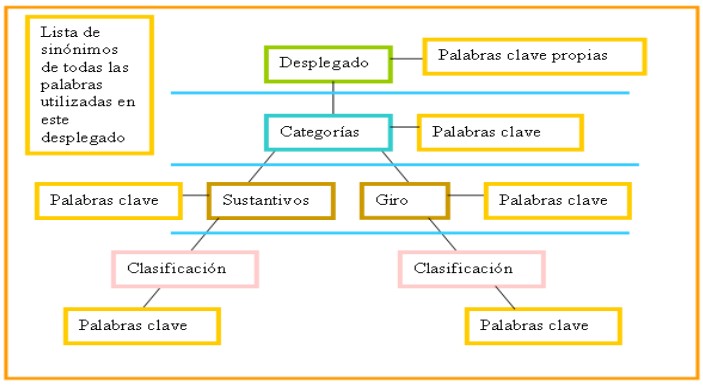
Los procedimientos de búsqueda van enfocados a encontrar negocios. Se pueden encontrar gracias a la herencia de palabras –resultado de las relaciones semánticas entre los elementos.

Mecanismo de búsqueda

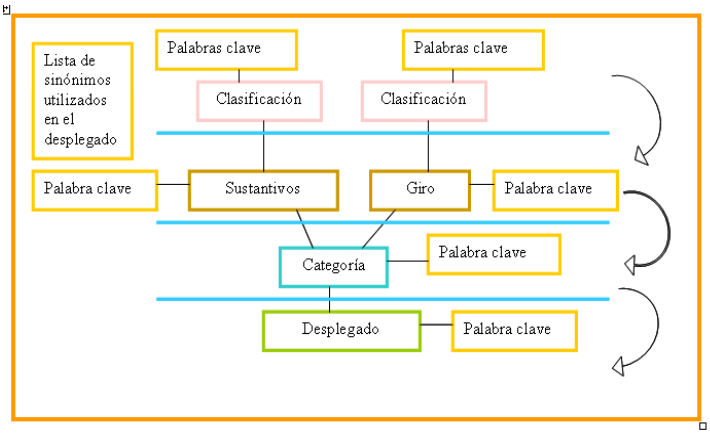
Ya se ha dicho que el sistema funciona por medio de la herencia de palabras que puede establecerse entre sus elementos.

Herencia de palabras entre los elementos del sistema:

- Los elementos del sistema se encuentran en distintos niveles:
    1. **Nivel 1: Capa externa: Resultados:**
    2. **Nivel 2: Capa media: Categorías**
    3. **Nivel 3: Capa interna: Sustantivos, giros**
    4. **Nivel 4: Capa madre: Clasificaciones**
    5. **Aditamentos: Palabras clave y sinónimos**

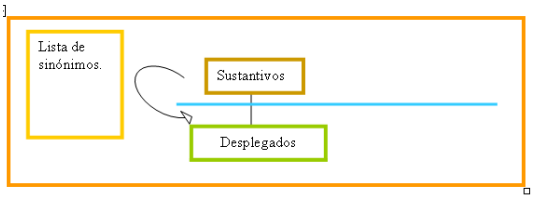


- Cada nivel hereda todo su contenido al siguiente:

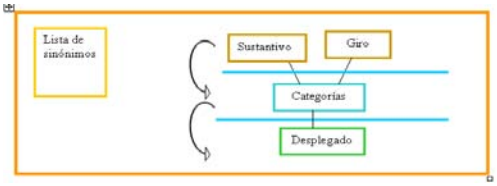


Esto significa que en una búsqueda la herencia de palabras hace más sencillo el proceso de encontrar desplegados. La herencia funciona:

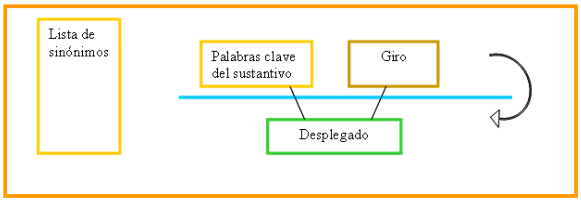
- Por medio de un sustantivo y los sinónimos que deriven de él:



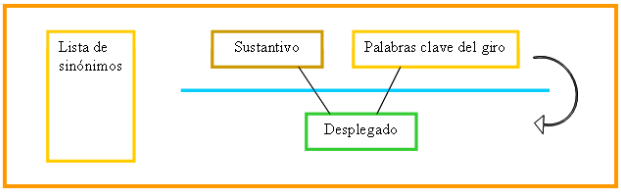
- Por medio de la combinación de un sustantivo y un giro (categoría) y los sinónimos de éstos:



- Combinación de palabras clave del sustantivo con el giro y los sinónimos de éstos:

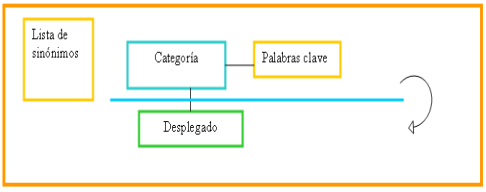


- Combinación de la palabra clave del giro con el sustantivo y los sinónimos de éstos:

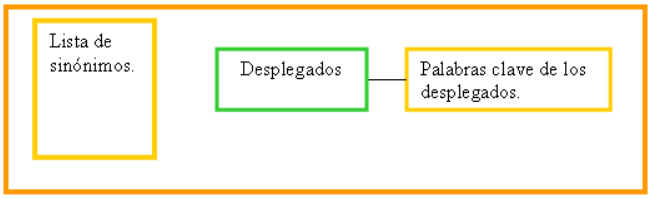


- Combinación de las palabras clave de los sustantivos y el giro y sus sinónimos:

- Por medio de las palabras clave de la categoría y sus sinónimo



- Por medio de las palabras clave propias de los desplegados y sius sinónimos:



La herencia de palabras para encontrar un desplegado no funcionará si solamente se pone un giro y las palabras derivadas de él

Bibliografía

Adam, Ch., y Tannery P. *Ouvres de Descartes*. Paris: J. Vrin, 1974

Borges, Jorge Luis. "El idioma analítico de John Wilkins". *Otras Inquisiciones,* Madrid: Alianza Editorial, 1997.

Codina, L., y Domènech, M. "Sobre los elementos a considerar en la representación del conocimiento de cara a la recuperación de información (el punto de vista cognitivo)". *La terminología científico-técnica*. Barcelona: IULATERM, 2001

Compañía Telefónica Mexicana. *Directorio telefónico de la Ciudad de México, año de 1891*. México: Condumex, 1991

Coseriu, Eugenio. *Introducción a la lingüística.* Madrid: Gredos, 1986.

Cuenca, M. J., y Hilferthy, J. *Introducción a la lingüística cognitiva*. Barcelona: Ariel, 1999.

Fillmore, C. "Frame semantics". *Linguistics in the morning calm*. Seul: Hanshin Publishing Co., 1982.

Focault, Michel. *Las palabras y las cosas*. México: Siglo XXI, 1967

Geckeler, H. *Semántica estructural y teoría del campo léxico*. Madrid: Editorial Gredos, 1976.

Kleiber, G. *La Semántica de los prototipos, categoría y sentido léxico*. Madrid: Visor Libros, 1990.

Lakoff, G. *Women, fire, and dangerous things, what categories reveal about the mind*. Chicago: The University of Chicago Press, 1990.

Lara, Luis Fernando. *Curso de lexicología*. México: El Colegio de México, 2006.

Lara, Luis Fernando. *Ensayos de teoría semántica, lengua natural y lenguajes científicos*. México: El Colegio de México, 2001.

Licklider, J.C.R. *Libraries of the future*. Cambridge, Mass: MIT Press, 1965.

Luque Durán, J. *Estudios de tipología lingüística*. Granada: Método Editorial, 1998.

Maldonado, R. "La Semántica en la gramática cognoscitiva*". Revista latina de pensamiento y lenguaje* vol. 1 (1993): 157-181 p.

Martí, M.A. *Tecnologías del lenguaje*. Barcelona: Editorial UOC, 2003.

Martí, M.A. *Lexicografía computacional y semántica*. Barcelona: Edicions Universitat de Barcelona, 2003.

Mattelart, A. *Historia de la sociedad de la información 1ª ed*. Barcelona: Bolsillo Paidos, 2007.

Mitkov, R. *The oxford handbook of computational linguistics*. New York: Oxford University Press, 2003.

Rosch, Eleanor. "Natural categories". *Cognitive Psychology* vol. 4, no. 3 (1976): 328-350 p.

Salton, G. "Automatic text analysis". *Science* vol. 168, no. 392 (1970): 335-343 p.

Swanson, D.R. "Searching natural language text by computer". *Science* vol. 132, no.3434 (1960): 1099-1104 p.

Taylor, J. *Linguistic categorization*. New York: Oxford University Press, 2003.

Téllez Valdez. *Derecho informático 3ª ed*. México: McGraw-Hill Interamericana Editores, 2004.

Wierzbicka, Anna. "Universal semantic primitives as a basis for lexical semantics". *Folia Linguistica: Acta Societatis Linguisticae Europaeae*, vol. 29, no. 1-2 (1995): 149-169 p.

Wittgenstein. L. *Investigaciones filosóficas*. México: Instituto de Investigaciones Filosóficas, UNAM, 1988.

Yus Ramos, F. *Ciberpragmática 2.0: nuevos usos del lenguaje en Internet.* Madrid: Ariel, 2010.

Fuentes electrónicas

Referencias en línea

Adame, Goddard, Lourdes. "Sección Amarilla: un impulsor estratégico de la telefonía comercial en México" [en línea]. *Instituto Mexicano de Teleservicios núm. 22*, 2008 http://www.imt.com.mx/rcforum/22/seccionamarilla.php

Blanco Cedón, Fernando. "Filosofía del lenguaje de Wittgenstein y el lenguaje de los científicos" [en línea]. *Canela Cuadernos*, *vol. XII*, 2000 http://www.canela.org.es/cuadernoscanela/canelapdf/cc12blanco.pdf

Cámara de la Fuente, Lidia. "La representación lingüística del conocimiento y su relevancia en la ingeniería lingüística" [en línea]. *Hipertext.net*, *núm. 2*, 2004 http://www.hipertext.net/web/pag224.htm.

Crespo García, Raquel. “Procesamiento del lenguaje natural” [en línea]. Departamento de ingeniería telemática. Universidad Carlos III de Madrid <volumen y fecha desconocidos> http://www.it.uc3m.es/rcrespo/docencia/irc/apuntes/Ingenieria%20Linguistica.pdf

Cuchi, Víctor. “Nacimiento de un mercado especial. El servicio telefónico en la Ciudad de México, 1881-1911” [en línea]. *Munich Personal RePEc Archive*, *núm. 1027*, 2007 http://mpra.ub.uni-muenchen.de/1027.pdf

Díez Orzas, P.L. «“Hiponimia y meronimia”. La relación de meronimia en los sustantivos del léxico español: contribución a la semántica computacional» [en línea]. *Estudios de lingüística del español*, *vol. 2*, 1999 http://elies.rediris.es/elies2/cap65.htm

Hilera, J. *“*Aplicación de técnicas de ingeniería lLingüística en sistemas de e-learning basados en objetos de aprendizaje” [en línea]. *Memoria del II Simposio Pluridisciplinar sobre Diseño, Evaluación y Descripción de Contenidos Educativos Reutilizables (SPDECE)*, 2005 http://www.uoc.edu/symposia/spdece05/pdf/ID02.pdf

Listerri, J. “Lingüística y tecnologías del lenguaje” [en línea]. *Lynx, Panorámicas de Estudios Lingüísticos*, *vol. 2*, 2003 http://liceu.uab.es/~joaquim/publicacions/Llisterri_03_Linguistica_Tecnologias_Lenguaje.pdf

Listerri, J. y Martí M.A. “La ingeniería lingüística en la sociedad de la información” [en línea]. *Digithum*, *revista digital d’ humanitats*, *vol. 3*, 2001 http://uoc.es/humfil/articles/esp/IIisterri-marti/IIisterri-marti.html

Moreno Ortiz, A., “«Sentando las bases para la construcción de lexicones automatizados en Europa: La Conferencia de Marina di Grosseto». *Diseño e implementación de un lexicón computacional para lexicografía y traducción automática*” [en línea]. *Estudios de Lingüística del Español, vol. 9*, 2000 http://elies.rediris.es/elies9/2-2-1.htm

Rodríguez Perojo, K y Ronda León, R. “Web semántica: un nuevo enfoque para la organización y recuperación de la información en la web” [en línea]. *ACIMED, vol. 13*, 2005 http://hdl.handle.net/10760/7961

Valderrábanos, Antonio. “Tecnología de lenguaje natural: cómo cerrar la brecha digital” [en línea]. Anobium.es, 2009 www.bitext.com

Libros en línea

Barité, Roqueta, Mario. “Diccionario de organización del conocimiento, clasificación, indización, terminología” [en línea]. Universidad de la República Oriental del Uruguay, 2011 http://www.eubca.edu.uy/kod/espaniol/index.php

Lamarca Lapuente, M.J. “Hipertexto: e nuevo concepto de documento en la cultura de la imagen”. Tesis [en línea]. Universidad Complutense de Madrid, 2009. http://www.hipertexto.info/documentos/h_hipertex.htm

Moreno Quibén, N. “Comprensión y producción semántica léxica”. Tesis [en línea]. UCLM-CEU Talavera, 2008. http://www.quiben.org/32030/wp-content/uploads/2008/01/apuntes-semantica.pdf

Tolosa, Gabriel. “Introducción a la recuperación de la información, conceptos, modelos y algoritmos básicos”. Tesis [en línea]. Universidad Nacional de Luján, 2007 http://www.tyr.unlu.edu.ar/tallerIR/2008/docs/Introduccion-RI-v9f.pdf

Real Academia Española. “Diccionario de la Real Academia Española” [en línea]. 2011 http://www.rae.es/rae.html