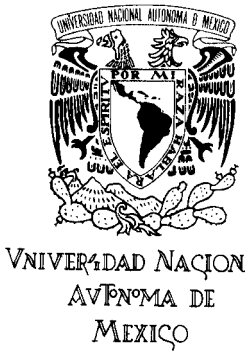

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"EXTRACCIÓN AUTOMÁTICA DE RELACIONES LÉXICO-SEMÁNTICAS A PARTIR DE TEXTOS ESPECIALIZADOS"

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

DOCTORA EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

P R E S E N T A :

OLGA LIDIA ACOSTA LÓPEZ

DIRECTOR DE LA TESIS: **DR. GERARDO E. SIERRA MARTÍNEZ**
INSTITUTO DE INGENIERÍA-UNAM

COMITÉ TUTORAL:
DRA. SOFÍA N. GALICIA HARO
FACULTAD DE CIENCIAS-UNAM
DR. MANUEL MONTES Y GÓMEZ
INSTITUTO NACIONAL DE ASTROFÍSICA, ÓPTICA Y ELECTRÓNICA

México, D.F. **FEBRERO 2013.**

Agradecimientos

Agradezco el apoyo en estos cuatro años de formación doctoral a mi tutor y también director de tesis, el doctor Gerardo Sierra Martínez, así como también a la doctora Sofía Natalia Galicia Haro y al doctor Manuel Montes y Gómez, miembros los tres de mi comité tutoral que contribuyeron con sus valiosas orientaciones y enseñanzas a concluir esta investigación.

A los doctores Francisco Hernández Quiróz y Christian Lemaitre León les agradezco la lectura minuciosa de este documento y las observaciones que sin duda contribuyeron a mejorar la calidad del escrito final.

A Amalia, Lulú, Diana y Cecilia, por la disposición y el apoyo incondicional brindado durante estos años de estudio y trabajo en el posgrado.

A mis amigos y colegas: Elias, Fernando, Verónica, Lizbeth, Olivia, Areli e Iván, así como también a todos los miembros del Grupo de Ingeniería Lingüística (GIL) con quienes he compartido momentos muy gratos y que han contribuido también al logro de esta meta.

A mis padres, hermanos y sobrinos, por ser la fuente de inspiración en mi vida.

A César, por confluir en tiempo y espacio para impulsarme a lograr esta meta tan importante en mi vida.

La presente investigación contó con el apoyo del Consejo Nacional de Ciencia y Tecnología (CONACYT) durante el período en el que se realizaron los estudios de posgrado (febrero 2008-febrero 2012). Esta investigación se llevó a cabo en el marco del proyecto CONACYT 82050: *Extracción de relaciones léxicas para dominios restringidos a partir de contextos definitorios en español*, y del proyecto PAPIIT de la UNAM, IN400312: *Análisis estilométrico para la detección de similitud textual*.

Abstract

This thesis is oriented to the automatic extraction of lexical sematic relations, particularly hyponymy-hypernymy relation from specialized texts. Although there are many works that have attempted to identify such relations, these have focused primarily on the analysis of English documents, and very few those analyze Spanish texts. Havind in mind this lack of works in Spanish, we propose here a method for recognizing and extracting hyponymy-hypernymy relation from specialised domains, useful for complementing learning methods of patterns and instances. For performing our method, we use a set of documents collected manually from MedLinePlus in Spanish, as well books of medicine.

In this regard, the current approaches of automatic learning of patterns consider the use of a set of *seed* instances –mainly, this set is manually provided-, in order to achieve a more efficient learning process based on reliable patterns, and then extract more realible instances. In brief, this thesis considers the following points:

1. The automatic extraction of a subset of instances of hyponymy-hypernymy relation implicit in specialized texts.
2. These relations are derived from textual fragments closely related to conceptual information, and which can serve as a *seed* set. In order to achieve this goal, we rely on a set of verbal patterns commonly used in analytic definitions. These patterns allow to identify and extract the term defined (hyponym) and its Genus term (hypernym).
3. Additionally, we take into account results from investigations in cognitive linguistic and psychology in order to obtain a great deal of valuable information of the set of hypernyms.

Keywords: Information extraction, lexical semantic relations, hypernymy, hyponymy, taxonomy.

Contenido

Índice de figuras

Índice de tablas

"*(...) By definition, an ontology is an explicit specification of a shared conceptualization. In essence, it is thus a view on how the world or a specific domain is structured as agreed upon by the members of a community. Assuming that we have perfect natural language processing tools for extracting knowledge from text, it is still questionable whether we will be able to actually learn an ontology from text as the conceptualization behind an ontology is typically assumed to be the result of an intentional process. Ontologies therefore can not be "learned" by machines in the strict sense of the word as they lack intention and purpose. Instead, ontology learning may only support an ontology engineer in defining their conceptualization of a particular part of the world, e.g. a technical domain, on the basis of empirical evidence derived from textual and other data."*

Ontology learning and population: Bridging the gap between text and knowledge, Buitelaar y Cimiano (2008).

Capítulo 1

Introducción

1.1 Punto de partida

El lenguaje natural escrito es uno de los principales medios de transferencia de conocimiento. A lo largo de la historia de la humanidad, generación tras generación, una cantidad significativa de conocimiento ha llegado hasta nosotros en la forma de lenguaje natural. Con el avance tecnológico adquirido hasta nuestros días, podemos acceder a enormes fuentes de información textual en formato digital. Aunado a lo anterior, el crecimiento exponencial de la información disponible en la Web ha generado la acumulación de vastas cantidades de textos relacionados con infinidad de temáticas.

El escenario anterior inevitablemente nos ha conducido a una situación en donde contamos con más información de la que somos capaces de leer, lo cual hace difícil muchas veces acceder eficazmente a ella, con miras sobre todo a explotarla para un fin específico.

Hoy en día resulta de gran interés extraer el conocimiento que subyace en enormes cantidades de textos especializados. La tarea de obtener este conocimiento manualmente, tan solo para un dominio específico, implica una revisión exhaustiva que consume demasiado tiempo y recursos, por ello la necesidad de buscar procesos alternativos (semi-)automatizados factibles que logren extraer la estructura del conocimiento implícito.

Uno de los enfoques comúnmente empleados para extraer conocimiento implícito de información textual viene de la semántica léxica, vista desde una perspectiva estructural (Croft y Cruse, 2004; Murphy, 2003). Si asumimos que las palabras denotan conceptos, la semántica léxica plantea la organización de un espacio conceptual por medio de relaciones léxico-semánticas de hiponimia-hiperonimia: *oftalmólogo-médico*, sinonimia: *inflamación-hinchazón*, meronimia-holonimia: *llanta-automóvil*, antonimia: *alto-bajo* y troponimia: *trotar-correr* (Fellbaum, 1998). Así, para efectos prácticos, el término relación léxico-semántica se entiende como aquellas relaciones que se establecen entre

los conceptos denotados por las palabras (Buitelaar *et al.*, 2005).

En este trabajo asumimos la postura anterior de que las palabras o *unidades léxicas* denotan conceptos y, debido a esto, consideramos necesario proporcionar una noción de concepto. En este sentido, asumimos una perspectiva cognitiva debido a que este campo de conocimiento ha contribuido con avances muy importantes en materia de conceptualización y se enfoca también en la comprensión de cómo el ser humano procesa el lenguaje. Por tanto, si logramos una comprensión más profunda de este fenómeno podríamos generar mejores modelos del lenguaje, lo que a su vez se traduciría en el desarrollo de aplicaciones más eficientes. Desde una perspectiva cognitiva, entonces, los conceptos reflejan la forma en que dividimos el mundo en clases y mucho de lo que aprendemos, comunicamos y razonamos incluye relaciones entre dichas clases. Así, si asumimos los conceptos como representaciones mentales de clases, estos cumplen una función muy importante, de acuerdo con Rosch (1978): la de promover la *economía cognitiva*.

La relación de hiponimia-hiperonimia ha sido una de las más trabajadas en la extracción de conocimiento tomando como fuente definiciones de diccionarios (Wilks *et al.*, 1996) y, más recientemente, de corpus de lengua general (Pantel y Pennacchiotti, 2006; Ortega *et al.*, 2007; Zhang *et al.*, 2011) o de dominio específico (Cimiano *et al.*, 2004; Acosta, 2009). Muchos diccionarios representan el significado de sus palabras mediante un *genus* y una o más *differentiae* (definición analítica). El *genus,* o también denominado hiperónimo, representa la clase o categoría a la que pertenece el término definido y la *differentia* lo distingue de otros términos pertenecientes a la misma clase. En el caso de corpus de dominio específico, las relaciones de hiponimia-hiperonimia pueden encontrarse implícitas en fragmentos lo más cercano a definiciones analíticas de conceptos: *la conjuntivitis se define como una inflamación de la conjuntiva del ojo*, o en otro tipo de expresiones, por ejemplo: *medicamentos tales como paracetamol y aspirina*. Dado lo anterior, la tarea de extracción de relaciones léxico-semánticas radica en la identificación de los patrones más característicos y confiables de la relación para con ellos extraer las instancias más confiables.

Los métodos propuestos a la fecha para la extracción de relaciones de hiponimia-hiperonimia se basan fundamentalmente en tres enfoques: correspondencia de patrones (Hearst, 1992; Acosta, 2009), distribucional (Pereira *et al.*, 1993; Faure y Nédellec, 1998; Caraballo, 1999) y subsunción de conceptos (Cimiano *et al.*, 2004). Además, debido a lo costoso, en tiempo y esfuerzo, que resulta la tarea de identificar manualmente los patrones más confiables de una relación en grandes repositorios de información textual se han explorado enfoques de aprendizaje automático para *aprender* los patrones más característicos (Snow *et al.*, 2005; Pantel y Pennacchiotti, 2006; Ortega *et al.*, 2007).

Los diferentes enfoques propuestos hasta ahora poseen ventajas y desventajas. Por un lado, entre las ventajas que ofrecen los métodos basados en correspondencia de patrones, como el propuesto por Hearst, se encuentra la precisión de los patrones identificados manualmente, sin embargo, algunos investigadores afirman que éstos representan una proporción muy pequeña, además de que la extracción manual de nuevos patrones consume demasiado tiempo y esfuerzo, por ello la necesidad de *aprenderlos* de forma automática. El aprendizaje automático de patrones e instancias constituye un avance muy importante, sin embargo, una de las desventajas principales es que necesita de una gran cantidad de información textual para enfrentar el problema de la *escasez de datos* (data sparseness), además de requerir de un conjunto de instancias *semilla* que en la mayoría de los casos se proporciona manualmente o con la ayuda de fuentes externas. Por otro lado, los enfoques distribucionales también requieren de una enorme cantidad de información textual para generar buenos resultados. Finalmente, los resultados obtenidos con el enfoque de subsunción de conceptos no han generado resultados relevantes.

A partir de lo expuesto anteriormente, a grandes rasgos, en este trabajo proponemos una metodología para la extracción automática de un subconjunto de instancias de la relación de hiponimia-hiperonimia de fragmentos textuales cercanamente relacionados con información conceptual conforme a la propuesta planteada por Sierra *et al.*, 2008 y Aguilar *et al.*, (2010). Para el caso de los métodos de aprendizaje automático, este subconjunto de instancias extraídas de información lo más cercana a

definiciones analíticas eliminaría la subjetividad en la selección del conjunto *semilla* y garantizaría su presencia en la fuente textual. Aunado a lo anterior y soportados por resultados de investigaciones en psicología cognitiva y lingüística cognitiva, consideramos que el conjunto de hiperónimos más frecuente puede ser de utilidad en un proceso de filtrado de información no relevante, así como también para obtener más hipónimos, cuya estructura *hiperónimo + modificador adjetivo* no estaría supeditada a que se aprendiera automáticamente el patrón, dado que asumimos *a priori* su importancia como reveladores de perspectivas de división relevantes del hiperónimo.

1.2 Estado del arte

1.2.1 Extracción conceptual

De acuerdo con Buitelaar *et al.* (2005) la extracción de conceptos de fuentes textuales es una tarea controvertida porque no resulta todavía claro lo que son. Desde su perspectiva, la formación o inducción de un concepto debe proporcionar los siguientes elementos:

- Una definición intensional
- Un conjunto de instancias, es decir, su extensión
- Una unidad léxica para denotarlo.

Los dos primeros elementos anteriores proporcionan el marco para explorar el estado del arte en la extracción conceptual, siendo en este punto de especial interés el primero, es decir, el intensional, debido a que en la presente investigación buscamos extraer fragmentos de texto donde se proporcione información de interés para un concepto específico. De acuerdo con Buitelaar *et al.* (2005) algunas investigaciones han enfatizado el aspecto extensional, es decir, localizar el mayor número de instancias de un concepto (Etzioni *et al.*, 2004; Evans, 2003). Por otro lado, la explotación del aspecto intensional de un concepto, es decir, la extracción de definiciones formales e informales ha sido más explotado dentro del marco del trabajo terminológico (Alarcón, 2009). Sin embargo, dentro de la tarea de extracción de información resulta también importante porque la información implícita en definiciones es útil para configurar un significado más preciso y completo de un concepto.

Extracción de conocimiento intensional

La siguiente revisión del estado del arte en cuanto a extracción conceptual está basada en el trabajo de Alarcón (2009), quien realizó un análisis muy completo y exhaustivo del tipo de metodologías existentes sobre extracción automática de contextos definitorios[1] para varios lenguajes.

Punto de partida

Uno de los puntos centrales de todas las metodologías propuestas es el uso de patrones que se usan comúnmente en definiciones. Dicho conjunto de patrones surge generalmente del análisis del comportamiento de estos fragmentos textuales en el contexto de un dominio específico. Derivado de este comportamiento se logra determinar la presencia regular de claves sintácticas y tipográficas que podrían ser de utilidad para la extracción automática de estos fragmentos relevantes[2] de información sobre un término. Por un lado, dentro de los patrones tipográficos encontramos signos de puntuación que vinculan el término con la definición, así como también aquellos mecanismos que son útiles para resaltar visualmente alguna información importante. Por otro lado, los patrones sintácticos están constituidos generalmente por verbos que introducen una predicación, que con frecuencia es una definición, para un término específico. A continuación, en la tabla 1 se muestra un resumen extraído y modificado de Alarcón (2009) sobre los patrones considerados en las diferentes investigaciones que se han realizado sobre el tema de extracción de información conceptual.

Como se observa de la tabla 1, la extracción de definiciones se basa prácticamente en patrones verbales, sean simples verbos (por ejemplo, *definir*, *denominar*) o vinculados con otros elementos (por ejemplo, *conocido como*). Existen casos, como el sistema Definder, donde se consideran patrones tales como *is the term for,* o en MOP que utiliza palabras simples como *Word* o *Term.* En lo que respecta a patrones tipográficos, lo más frecuente son los signos de puntuación, que en gran medida corresponden al uso de paréntesis.

[1] Un contexto definitorio es un fragmento de texto donde se define un término.

[2] Por *relevante* entendemos aquella información que es de interés o sirve para los propósitos de un dominio específico.

Finalmente, es importante mencionar que este tipo de metodologías se han desarrollado para el inglés y el francés. La implementación a otras lenguas tales como el portugués o español es relativamente reciente. Concretamente para el caso del español, las únicas referencias que Alarcón (2009) documenta son los casos de Sánchez y Márquez (2005) y Corpógrafo (Sarmento *et al.*, 2006).

Tabla 1. Ejemplos de patrones utilizados en la extracción de información conceptual. Extraído y modificado de Alarcón (2009).

Referencia[3]	**Lengua**	**Tipo de patrón**	**Ejemplos**
1. Rebeyrolle y Tanguy (2000)	Francés	Verbal	Défin$ Definir * comme
2. Definder (Muresan y Klavans, 2002)	Inglés	Signos Verbal Frases	Paréntesis, guiones Is called Is the term for
3. Malaisé (2005)	Francés	Signos Verbal Marcadores	Paréntesis Dénnomer Nom , c'est-à-dire
4. Sánchez y Márquez (2005)	Español	Verbal	Entenderse
5. Storrer y Wellinghoff (2006)	Alemán	Verbal	Definieren als
6. MOP (Rodríguez, 2004)	Inglés	Signos Verbal Marcadores	Comillas Called, known as Word, term
7. Saggion (2004)	Inglés	Verbal	TERM is a, such as TERM
8. LT4eL (Monachesi, 2007)	Alemán, búlgaro, checo, holandés, inglés, polaco, maltés, portugués, rumano	Signos Tipografía Verbal	Paréntesis Viñetas IS-A, defined as, Called
9. GlossExtractor (Navigli y Velardi, 2007)[4]	Inglés	Verbal	TERM refers to TERM is a kind of
10. Corpógrafo (Sarmento et al., 2006)	Alemán, español, inglés, italiano, francés, portugués	Verbal	TERM is a * that Are known as TERM

[3] Por facilidad de referencia, el número que precede a la referencia es el utilizado también en la tabla de resultados de evaluación.

[4] *GlossExtractor* es una aplicación Web que recibe como entrada el resultado de una aplicación de extracción terminológica denominado *TermExtractor*, o bien una terminología proporcionada por el usuario.

Filtrado de información no relevante

Como en todos los procesos de extracción automática de información existe siempre la posibilidad de obtener *ruido* en los resultados. La extracción de información conceptual no es la excepción debido a que en el caso de algunos verbos como *ser*, se usan en una gran variedad de contextos. Dado lo anterior, en la mayoría de las metodologías se han implementado filtros para eliminar el *ruido* presente en los datos. Estos filtros van desde la restricción de tiempos verbales que recuperan mayor *ruido* hasta la aplicación de aprendizaje máquina para determinar los elementos causantes del *ruido*. La tabla 2 concentra la información sobre las estrategias asumidas en cada una de las metodologías revisadas por Alarcón (2009).

Tabla 2. Filtros implementados en cada metodología.

Referencia	**Filtro**
Rebeyrolle y Tanguy (2000)	Inclusión de ventanas y restricción del tipo de elementos en este conjunto: *definir * como.*
Storrer y Wellinghoff (2006)	Consideración de patrones más específicos, por ejemplo: *se entiende por, se define como,* etc.
Rodríguez (2005)	Aplicación de aprendizaje máquina para encontrar patrones recurrentes en contextos no relevantes.
LT4eL (Monachesi, 2007)	Algoritmos genéticos para asignar un peso mayor a patrones definitorios que presenten mejores resultados.
GlossExtractor (Navigli y Velardi, 2007)	Expresiones regulares identificadas mediante algoritmos de aprendizaje máquina para extraer definiciones analíticas.
Saggion(2004)	Filtrado mediante términos secundarios más relacionados con el término de búsqueda.

Metodologías de evaluación

De acuerdo con Alarcón (2009), el análisis contrastivo de metodologías de evaluación relacionadas con la extracción automática de información conceptual no es una tarea trivial debido a que existen problemas que hacen de este proceso uno de los menos estandarizados. El problema principal es decidir qué es y qué no es una definición. Aunado a lo anterior, surge el problema de qué es una buena definición, y esto sin duda repercute en los criterios a considerar para construir un *golden standard* de tal suerte que se

cuente con una buena fuente de definiciones para procesos de evaluación en la extracción automática.

En el mismo orden de ideas y para enfrentar la situación anteriormente descrita, lo que se deduce a partir de la revisión de metodologías de los trabajos sobre extracción conceptual es que todos se basan en un conjunto de definiciones recolectado de las fuentes textuales analizadas, el cual constituye un *golden standard* para objetivos de evaluación de los procesos de extracción automática. Lo que en todo caso diferencia unas metodologías de otras es la forma en que fueron seleccionadas estas definiciones. Por ejemplo, algunos involucraron más de una persona para determinar las buenas definiciones mediante un acuerdo inter-anotador (Muresan y Klavans, 2002). Otros simplemente se basaron en su competencia lingüística para tomar esta decisión (Rebeyrolle y Tanguy, 2000; Malaisé, 2005; Navigli y Velardi, 2007, Sánchez y Márquez, 2005; Rodríguez, 2005).

Resultados de evaluación

Los resultados obtenidos en la aplicación de las diferentes metodologías mencionadas en el estado del arte sobre extracción automática de información conceptual se presentan en la figura 1, extraída sin cambios también de Alarcón (2009). La figura 1 resume los resultados en términos de precisión[5] (P), cobertura (C) y medida F (F1 y F2). Alarcón menciona que uno de los criterios considerados para la construcción de esta tabla fue que se basara en la evaluación del proceso de extracción de contextos en general. Algunos trabajos no presentaron resultados de este tipo, por lo cual este rubro se encuentra ausente en algunos casos.

Como puede observarse de la figura1, la mayoría de las metodologías considera dominios especializados, salvo el caso del trabajo de Saggion (2004) que se llevó a cabo en noticias, que se considera una fuente más cercana a lengua general. El tamaño de la fuente textual muestra también diferencias notables, y se describe en la tabla como número de oraciones (o), palabras (p) o texto (t).

[5] Precisión es la fracción de instancias recuperadas que son relevantes, mientras que cobertura es la fracción de instancias relevantes que se recuperan.

Rf.	Id.	Área	Tamaño	P	C	F1 / F2
1	fr	Geomorfología	275,000 p			
		Ing.		-	-	-
		conocimiento	230,000 p			
		Empresariales	205,000 p			
		Enciclopedia	215,000 p			
2	in	Medicina	-	86.95%	75.47%	-
3	fr	Dietética y nutrición	480,000 p	55%	39.3%	-
4	es	Derecho	-	97.44%	100%	-
5	al	Tecnologías texto	103,805 p	34%	70%	-
6	in	Sociología	5,581 o	0.97	0.79	F1 0.87
		Histología	5,146 o	0.94	0.81	F1 0.87
7	in	Noticias	1,000,000 t	-	-	F1 0.23
8	al	Derecho	237,935 o	48.6%	-	-
	bu	Cómputo	76,800 p	22.5%	8.9%	F2 11.1%
	ch	" "	90,000 p	22.3%	46%	F2 33.9%
	ho	" "	14 t	-	-	-
	pol	" "	83,200 p	23.3%	32%	F2 28.4%
	por	" "	274,000 p	0.14	0.86	F2 0.66%
	ru	" "	700,000 p	-	-	-
9	in	Economía	1,000 o	0.87	0.86	F1 0.86
		Medicina	250 o			
10	in	Medicina	80,295 p	-	-	-
	por	" "	21,667 p			

Figura 1. Resultados de la extracción conceptual.
Extraído de Alarcón (2009).

Existen también diferencias significativas en los resultados de una metodología a otra, lo que se explica por la diferencia en los patrones considerados, los filtros implementados para la eliminación de ruido en los resultados y la cobertura de su análisis. Un ejemplo de esto último es el trabajo de Sánchez y Márquez (2005) que se enfoca en un verbo específico. Por otro lado, el caso de Saggion (2004) reporta una puntuación F de 0.23, que por lo menos supera un desempeño promedio en los resultados logrados en la competencia TREC QA 2003 (0.192), sin embargo, se queda por debajo del desempeño más alto de 0.55. Por su parte Rodríguez (2004) señala que sus indicadores de precisión y recuperación son comparables con aquellos obtenidos con Definder, sin embargo, desde el punto de vista de Alarcón (2009), existen diferencias en cuanto a la forma de conformación del *golden standard* de definiciones, situación que resta conveniencia a la comparación.

Finalmente, la diferencia en resultados del sistema LT4eL, de acuerdo con Alarcón (2009), se explica porque en el caso de las lenguas eslavas, como el búlgaro, las definiciones se enuncian en más de una oración y la metodología general en el proyecto sólo contemplaba una oración como extensión máxima de un contexto definitorio, es por ello que se obtienen indicadores más bajos de precisión y cobertura en este tipo de lenguas.

Ecode: Sistema de extracción de contextos definitorios para el español

La metodología para la extracción de contextos definitorios para el español está basada en reglas lingüísticas y contempla un conjunto de patrones verbales que permiten la identificación de contextos donde se puede encontrar información importante respecto a un término. Estos patrones tienen como núcleo un verbo y, en algunos casos, otros elementos pueden estar ligados al mismo, por ejemplo, en el caso del verbo *definir* podemos encontrar construcciones donde se liga con el adverbio *como* en la expresión *definir como.* La figura 2 muestra los verbos considerados en esta metodología. La determinación de este conjunto de verbos y sus estructuras es producto de una serie de trabajos de investigación llevados a cabo por el Grupo de Ingeniería Lingüística (GIL) de la UNAM.

Ecode recibe como entrada un corpus etiquetado con partes de la oración y realiza una extracción de contextos vía una gramática de patrones verbales donde se establecen datos específicos relacionados con cada verbo utilizado en definiciones. Además, otra característica del sistema es que considera restricciones verbales de tiempo y persona gramatical, así como también distancias entre el verbo y un posible nexo. Otra particularidad más del sistema es que contempla la extracción de varios tipos de contextos defnitorios: analíticos, sinonímicos, extensionales y funcionales, que identifica a partir de los patrones verbales considerados.

La salida del sistema Ecode es un conjunto de contextos definitorios clasificados de acuerdo a si son analíticos, extensionales, sinonímicos y funcionales. Además, ofrece un mecanismo para ordenar por relevancia estos contextos y la identificación del término, así como también de la definición.

Tipo	Lema	Raíz	Nexo
Analítico	ser	(es|son)	determinante
	caracterizar	caracteriz	como, por
	concebir	conc(e|i)b	como
	considerar	consider	como
	describir	describ	como
	definir	defin	como
	entender	ent(ie|e)nd	como
	conocer	conoc	como
	denominar	denomin	∅, como
	llamar	llam	∅, como
	nombrar	nombr	∅, como
Extensional	comprender	comprend	∅
	contener	cont(ien|en|uv)	∅
	incluir	inclu(í|i|y)	∅
	integrar	integr	∅
	constar	const(a|e|ó)	de
	contar	c(ue|o)nt(a|e|á|é|ó)	con
	formar	form	de, por
	componer	comp(on|us|uest)	de, por
	constituir	constit	de, por
Funcional	permitir	permit	∅
	encargar	encarg	de
	consistir	consist	en
	funcionar	funcion	como, para
	ocupar	ocup	como, para
	servir	s(i|e)rv	como, en, para
	usar	us	como, en, para
	emplear	emple	como, en, para
	utilizar	util	como, en, para
Sinonímico	conocer	conoc	también
	denominar	denomin	también
	llamar	llam	también
	nombrar	nombr	también

Figura 2. Modelo de verbos usados en definiciones, raíces y nexos en la gramática de patrones verbales de Ecode. Extraída de Alarcón (2009).

Identificación de los elementos principales: término y definición

Ecode realiza un proceso de identificación del término y la definición considerando las posiciones que tienden a seguir de acuerdo con el verbo que los introduce. A esta configuración de posiciones se le denomina patrón contextual (Alarcón, 2009). Este proceso se realiza posterior a la etapa de filtrado de contextos no relevantes. Además, esta fase también constituye un filtro porque si no se logra extraer un término y una definición con una estructura específica, el contexto definitorio se excluye del conjunto de buenos candidatos. La figura 3 presenta un conjunto de patrones contextuales que se tomaron en cuenta en este sistema para el caso del verbo *definir*.

Patrón contextual	Ejemplo
T + VD + NX + D	T se define como D
D + VD + NX + T	D se define como T
VD + T + NX + D	se define T como D
VD + NX + T + D	se define como T D
PPR + T + VD + NX + D	generalmente T se define como D
T + PPR + VD + NX + D	T generalmente se define como D
VD + PPR + T + NX + D	se define generalmente T como D
VD + T + PPR + NX + D	se define T generalmente como D

Figura 3. Patrones contextuales del verbo *definir*.
Extraído de Alarcón (2009).

Dado lo anterior, el proceso de identificar término y definición consiste en fijar de entrada el patrón verbal y después determinar qué elementos forman parte del término y cuáles de la definición. El proceso principal para identificar el término y la definición se lleva a cabo mediante un árbol de decisiones basado en la gramática de patrones verbales, específicamente en los patrones contextuales, como el presentado en la figura 3, para el caso del verbo *definir*. Además, aunado al patrón contextual, también se basa en algunas etiquetas gramaticales que dan cuenta de la estructura que debe tener un término y su definición.

Evaluación del sistema Ecode

La precisión y cobertura global del sistema Ecode se evaluó sobre un subconjunto del Corpus Técnico del IULA en español. Con base en el objetivo de la evaluación respecto a un sistema basado en patrones cuyo núcleo es un verbo, se conformó este subconjunto de textos a partir de buscar los lemas de los verbos contenidos en la gramática de patrones. De los resultados obtenidos para cada lema se seleccionaron 250 ocurrencias. Los resultados globales se presentan en la tabla 3. De acuerdo con Alarcón (2009) estos resultados globales no incluyen restricciones de algún tipo.

Tabla 3. Desempeño del sistema Ecode.

Restricción	Precisión	Cobertura
Sin restricción	53%	79%
Tiempo y persona	56%	73%
Distancia entre verbo y nexo	56%	72%
Patrón contextual	57%	70%

En la siguiente tabla se muestran los indicadores de precisión y cobertura por tipo de definición. Como se puede observar, las definiciones sinonímicas son las que presentan una mayor precisión seguida de las definiciones analíticas. Con una cobertura para ambas por arriba del 80%. Por otro lado, las definiciones extensionales y funcionales no alcanzan una precisión del 50%, sin embargo, los niveles de cobertura son altos.

Tabla 4. Desempeño de Ecode por tipo de definición.

Tipo de contexto definitorio	Precisión	Cobertura
Analítica	58%	83%
Extensional	48%	77%
Funcional	45%	83%
Sinonímica	76%	85%

Por su parte, la aplicación de restricciones como tiempo y persona gramatical en los verbos generó indicadores globales del 56% y 73% para precisión y cobertura, respectivamente. Otra restricción adicional relacionada con la distancia entre el verbo y su posible nexo en una ventana de 15 palabras no logra un incremento en la precisión, que se queda en un 56% respecto al resultado con la primera restricción, y la cobertura baja un punto porcentual (72%). Finalmente, en lo que respecta a los patrones contextuales, la precisión aumenta hasta un 57%, sin embargo, la cobertura desciende hasta un 70%.

1.2.2 Extracción de relaciones léxico-semánticas

Existe una buena cantidad de trabajos en torno a la extracción de relaciones de hiponimia-hiperonimia entre términos/conceptos enfocados principalmente

al procesamiento de textos en inglés, los cuales pueden clasificarse en tres enfoques principales (Cimiano *et al.*, 2004; Ryu y Choi, 2005):

1. Enfoque basado en la correspondencia de patrones léxico-sintácticos.
2. Enfoque de agrupamiento (clustering) basado en la distribución del contexto en un corpus.
3. Enfoque de subsunción basado en corpus.

El primer enfoque consiste en la obtención de patrones léxico-sintácticos que caractericen una relación léxico-semántica de interés. En el marco de este paradigma, el trabajo que se considera pionero es el de Hearst (1992). Dicho trabajo consistió en un algoritmo de adquisición de patrones que caracterizaran una relación específica a partir de proporcionar un conjunto de instancias de la relación. Una vez localizados nuevos patrones, éstos se utilizan en una etapa *bootstrapping*[6] para localizar nuevas instancias. Algunos trabajos posteriores se han basado en el método *bootstrapping* propuesto para extraer relaciones de meronimia-holonimia (Berland y Charniak, 1999) e hiponimia-hiperonimia (Pantel y Pennacchiotti, 2006; Ortega *et al.*, 2007). En estos trabajos se ha considerado importante la obtención de una mayor cantidad de patrones que caractericen la relación de interés y para ello se parte de un conjunto de pares de términos relacionados mediante la relación objetivo. Otra de las ventajas de estos enfoques es que han introducido medidas de confiabilidad para evaluar la calidad de los patrones y de las instancias obtenidas a partir de los mismos, dicha fase no había sido considerada en el trabajo de Hearst. Finalmente, Acosta (2009) diseñó una gramática de restricciones para identificar estructuras sintácticas de los elementos más comunes en candidatos a definiciones analíticas para finalmente extraer término e hiperónimo.

Por otro lado, hay también esfuerzos enfocados en reducir el trabajo manual en la tarea de identificar patrones léxico-sintácticos que caractericen de manera precisa una relación léxica de interés utilizando, por ejemplo, técnicas de aprendizaje máquina. Los patrones extraídos se combinan después usando un algoritmo de aprendizaje supervisado para obtener un clasificador

[6] En este contexto, el término *bootstrapping* se refiere al proceso de extracción de una mayor cantidad de relaciones mediante un conjunto pequeño de patrones o instancias de la relación objetivo.

que tenga una precisión alta (Snow *et al.*, 2005). Este tipo de enfoques requiere de corpus etiquetados sintácticamente para extraer los patrones más característicos de la relación y, para el caso del inglés, existe una gran cantidad de recursos de este tipo (Brown, Penn Treebank, WSJ, etc.); sin embargo, para otros lenguajes, como el español, esta disponibilidad puede llegar a ser prácticamente nula.

Por su parte, el enfoque basado en agrupamiento considera la distribución contextual derivada de un corpus, conocida como hipótesis distribucional de Harris (Harris, 1970). Esta hipótesis sostiene que las palabras que tienen un significado similar ocurren en contextos similares. En el marco de este enfoque, Pereira *et al.* (1993) proponen un método para agrupar palabras de acuerdo con su distribución contextual, concretamente tomando en cuenta contextos sintácticos. En el marco de este trabajo, las palabras son representadas por las distribuciones de frecuencias relativas de los contextos en los que aparecen, utilizando la entropía relativa entre estas distribuciones como medida de similitud para la agrupación. Por otro lado, Faure y Nédellec (1998) presentan un enfoque de agrupamiento de abajo hacia arriba (bottom-up) iterativo de nombres[7] en contextos similares. En cada paso, se agregan dos clases si la distancia es menor a un umbral establecido por el usuario. Esta distancia se define como la proporción de palabras núcleo común en los dos grupos considerando su frecuencia. Por último y también dentro de este paradigma, otro trabajo que es importante mencionar es aquel llevado a cabo por Caraballo (1999), quien utilizó este enfoque para generar jerarquías de nombres considerando datos de conjunciones y aposiciones de nombres recolectados de un corpus.

Finalmente, el enfoque de subsunción basado en corpus establece la generalidad relativa de dos términos a partir de considerar que un término t_1 es una subclase de t_2 si todos los contextos sintácticos en los que t_1 aparece son también compartidos por t_2. En el marco de este paradigma, Cimiano *et al.* (2004) proponen extraer del corpus las dependencias pseudo-sintácticas de cada término. Estas dependencias no son obtenidas de análisis sintácticos profundos, sino de métodos sintácticos superficiales (shallow parsing) que

[7] En este trabajo, se denomina *nombre* a la categoría de palabra *sustantivo*.

trabajan con expresiones regulares sobre un corpus con etiquetado gramatical. Entre las expresiones que consideran están: adjetivos, frases preposicionales (FPs), frases nominales[8] (FNs) en posición de sujeto u objeto, frases preposicionales que siguen a un verbo, etc. Una vez obtenidos los vectores de términos con sus rasgos correspondientes se aplica una medida para determinar un índice de los rasgos de t_1 incluidos en los rasgos de t_2. Desafortunadamente, los experimentos realizados con este método por Cimiano *et al.* (2004) reportan niveles de cobertura relativamente aceptables (27.83%), pero una precisión muy baja (0.92%).

En resumen, los enfoques basados en patrones para la extracción de relaciones léxico-semánticas consideran como punto de partida un conjunto *semilla* que se ha proporcionado manualmente en la mayoría de los casos. Este conjunto de pares de elementos debe mantener la relación de interés y con ello contribuir a localizar patrones que la expresen y, posteriormente, más instancias. No existen estudios previos respecto a los criterios para la selección de este conjunto inicial, por lo que es difícil, por el momento, determinar su influencia en la obtención de patrones e instancias. Aunado a lo anterior, consideramos que la selección de este conjunto implica necesariamente que el usuario de un sistema de extracción con estas características tenga conocimiento del dominio. Dada esta situación, desde nuestra perspectiva, este subconjunto de instancias debe ser extraído de forma automática de la fuente textual donde se pretende *aprender*. Otra desventaja importante se encuentra en la gran cantidad de datos requerido para enfrentar el problema de la *escasez de datos* (data sparseness) y con ello aumentar la probabilidad de encontrar diferentes formas de expresión de la relación. En este sentido, la enorme cantidad de recursos disponibles en la Web puede ofrecer una buena solución, sin embargo, creemos que es también necesario hacer énfasis en los tipos de fuentes textuales que se consideran para *aprender* si es que se pretende que el conocimiento adquirido alimente o ayude a construir recursos tales como tesauros, taxonomías u ontologías. Finalmente, los enfoques basados únicamente en información sintáctica no han logrado resultados

[8] En este contexto, una frase nominal es una cadena de palabras configurada en torno a un nombre (núcleo nominal).

sobresalientes, por lo que resulta de gran interés complementarlos con información adicional.

Por su parte, el enfoque basado en agrupamiento tiene la desventaja de generar grupos no etiquetados y donde resulta difícil asignar, de forma automática, etiquetas adecuadas a los grupos; además de requerir de corpus de gran tamaño para un mejor desempeño. A este respecto, como lo señalan Pantel y Pennacchiotti (2006), los enfoques basados en agrupamiento no supervisado generalmente no producen grupos coherentes para corpus menores a 100 millones de palabras, por lo que no resultan confiables para colecciones de textos más pequeños. Por último, el enfoque de subsunción de términos plantea una revisión exhaustiva de los contextos sintácticos de cada término para derivar la jerarquía de los mismos, lo que puede ser de gran utilidad para derivar información relevante respecto al hipónimo y el hiperónimo, sin embargo, los experimentos realizados hasta ahora no arrojan resultados sobresalientes.

1.3 Objetivos

Aunque nuestro objetivo no es la construcción de bases de conocimiento léxica, tesauros, taxonomías u ontologías como producto final, los resultados derivados de esta investigación serán de utilidad para soportar el mantenimiento o parte de la construcción de estos recursos. Sin embargo, para lograr lo anterior, es necesario enfatizar la importancia de contar con fuentes textuales adecuadas para la extracción de conocimiento. Como Buitelaar y Cimiano (2008) lo señalan, un aspecto importante que sigue aún sin consensuarse es precisamente el tipo de evidencia textual que debe considerarse cuando se modelan estos recursos. En este sentido, Cabré *et al.* (2000) plantean que el grado de especialización de un texto varía en un continuo desde un nivel bajo (prensa escrita y revistas de gran difusión), seguido por un nivel medio (artículos de divulgación científica, secciones técnicas de diarios y revistas), hasta un nivel alto (artículos científicos). Dado lo anterior y asumiendo la visión de ontologías de Gruber (1993), consideramos que el grado de especialización más adecuado para la extracción de conocimiento debe ser medio o alto, priorizando el nivel medio, porque es donde existe una probabilidad mayor de encontrar definiciones que reflejen, en

mayor o menor grado, un consenso respecto al conocimiento del dominio.

Considerando lo expresado en el párrafo anterior, el objetivo general de este trabajo de investigación consiste en desarrollar una metodología para la extracción automática de un subconjunto de relaciones léxico-semánticas de hiponimia-hiperonimia implícitas en candidatos a contextos definitorios extraídos de corpus de dominio específico. *Grosso modo*, la utilidad de este subconjunto podría ser la siguiente:

1. Conjunto *semilla* en métodos de aprendizaje automático de patrones e instancias. Este conjunto eliminaría el sesgo introducido mediante una selección subjetiva de pares de la relación y reduciría también el nivel de conocimiento del dominio requerido por parte del usuario debido a que las instancias se extraen del mismo corpus de donde se pretende *aprender*.
2. Conjunto de hiperónimos como filtro de información no relevante si se aplica un enfoque probabilístico, es decir, bajo este criterio los hiperónimos más frecuentes obtenidos de los candidatos a contextos definitorios tendrían una mayor probabilidad de ser verdaderos. Por tanto, los fragmentos candidatos que tengan este hiperónimo tendrían una probabilidad mayor de contener relaciones verdaderas.

Los objetivos específicos que soportan el objetivo general planteado anteriormente son los siguientes:

- Diseñar una gramática de expresiones regulares para una fase de análisis sintáctico superficial (chunking) que considere el comportamiento de los constituyentes más comunes de los contextos definitorios analíticos, así como también de sus patrones contextuales. En este sentido, se prioriza la precisión de los análisis sintácticos de definiciones realizados por Aguilar (2009) y se complementan con una mayor evidencia empírica derivada de un conjunto de definiciones obtenidas manualmente de Wikipedia. La gramática de patrones se construye manualmente en lugar de *aprenderse* automáticamente debido a que el aprendizaje requiere de análisis sintácticos más profundos que determinen la estructura de constituyentes de los elementos de una oración, y la precisión de estos analizadores, por lo

menos para el español, todavía se encuentra por debajo del 90% (Ballesteros *et al.*, 2010).

- Explorar filtros de información no relevante que permitan la obtención de los mejores candidatos a contextos definitorios analíticos y, en consecuencia, de las mejores instancias de la relación de hiponimia-hiperonimia.
- Extraer una mayor cantidad de información del conjunto de hiperónimos. Planteamos este objetivo basados en resultados derivados de investigaciones de las ciencias cognitivas, concretamente de la psicología y de la linguística.
- Explorar filtros de hipónimos no relevantes generados a partir de un hiperónimo más modificadores adjetivos.
- Explorar fuentes de hiperónimos alternas para hacer más robustos los procesos de filtrado de información no relevante.

1.4 Hipótesis

Dado el uso de la teoría clásica como una meta-teoría para formular conceptos en ciencia, las definiciones analíticas son comunes en textos especializados y siguen un patrón de comportamiento sintáctico regular. Tomando en cuenta esta regularidad sintáctica, es posible automatizar la obtención de un subconjunto de instancias confiable de la relación de hiponimia-hiperonimia para efectos de servir como conjunto *semilla* en enfoques de aprendizaje automático de patrones e instancias.

1.5 Aportaciones

En términos generales, la aportación al estado del arte actual es un programa, denominado ExtReLex, que a partir de un corpus de dominio específico extrae un conjunto de contextos definitorios analíticos candidatos y las relaciones de hiponimia-hiperonimia implícitas en estos fragmentos textuales. Además, extrae un conjunto de hipónimos derivados de los hiperónimos más frecuentes.

Las aportaciones específicas que están implícitas en el programa ExtReLex son las siguientes:

1. Un modelo de constituyentes para los elementos más comunes insertos

en definiciones analíticas. Estos constituyentes contemplan la estructura más común de términos, sinónimos de términos, patrones pragmáticos, hiperónimos y patrones verbales, estos últimos sirven como vínculos entre términos y definiciones.

2. Un subconjunto de instancias de la relación de hiponimia-hiperonimia implícita en textos especializados que se obtiene de forma automática y que puede ser usado como conjunto *semilla* en métodos de aprendizaje automático. Debido a que este subconjunto se puede extraer del mismo corpus de donde se pretenda *aprender*, esto garantiza su presencia.

3. Un método que considera heurísticas lingüísticas para el filtrado de hipónimos no relevantes generados a partir del hiperónimo. Este método supera los resultados de medidas de la teoría de la información, tales como la información mutua puntual, dada la composicionalidad que existe entre los hiperónimos y sus modificadores adjetivos. Basados en resultados de investigaciones de la psicología cognitiva y la lingüística cognitiva justificamos la importancia de extraer estas categorías subordinadas sin necesidad de supeditarlas al resultado de un proceso de *aprendizaje*.

4. Una comprensión más clara de los diferentes tipos de relaciones de hiponimia-hiperonimia localizados en fuentes textuales a la luz de la teoría clásica de conceptos y de los resultados de investigaciones en psicología cognitiva.

1.6 Pasos esenciales de la tesis

En este apartado describimos las decisiones iniciales más importantes para llevar a cabo nuestra investigación. Una de estas decisiones se relaciona con el tipo de fuente de información textual. Dado que nuestro trabajo toma como punto de partida información lo más cercana a lo conceptual, consideramos que la fuente más apropiada y confiable para localizarla son precisamente los dominios especializados. El énfasis en dominios restringidos y la definición *a priori* de la información de interés circunscribe esta investigación en el área de extracción de información. El enfoque inicial basado en patrones para tareas de extracción de información, en menor o mayor medida, continúa siendo

explotado en la actualidad. La regularidad en la expresión de información relevante mediante el lenguaje natural facilita su extracción. Un caso concreto del tipo anterior lo representan las definiciones analíticas, un esquema de definición de conceptos que refleja la adopción de la visión clásica como una meta-teoría para formular conceptos científicos.

1.6.1 Selección de la fuente de información

La creatividad en las formas de expresión de los seres humanos a través del lenguaje no parece tener límites, lo que sin duda representa una dificultad seria para los diferentes métodos de recuperación y extracción de información que se han propuesto a la fecha. Así, en el caso del lenguaje que utilizamos para interactuar con otros en un entorno informal o casual, sea éste hablado o escrito, nos enfrentamos a demasiada ambigüedad, vaguedad e información que queda muchas veces implícita. Dado lo anterior, si se pretende extraer información de fuentes de lengua general para construir, por ejemplo, una ontología, la brecha semántica entre el lenguaje natural y el conocimiento formalizado que se persigue como objetivo es mucho más amplia.

La selección de información lo más cercana a lo conceptual (contextos definitorios) como punto de partida para la extracción de relaciones léxico-semánticas se da a partir de considerar que los conceptos son el mecanismo por medio del cual dividimos el mundo en clases. De acuerdo con Rosch (1978), la función más importante de los conceptos es la de promover la *economía cognitiva*, es decir, proveer de máxima información con el menor esfuerzo cognitivo. Los conceptos están presentes en cualquier escenario de comunicación y nos salvan del manejo de enormes lexicones que harían la comunicación prácticamente imposible. Sin embargo, los dominios restringidos o especializados circunscriben su conocimiento al manejo de un conjunto de conceptos léxicos (su terminología) cuyas relaciones representan el conocimiento del dominio y pueden contribuir a conceptualizarlo.

Finalmente, esta tesis se enfoca en la extracción automática de relaciones léxico-semánticas de tipo hiponimia-hiperonimia a partir de fuentes textuales no estructuradas, lo que enmarca la investigación en el área de extracción de información. Las fuentes textuales que se consideran en este trabajo son textos de dominio específico debido a que estas fuentes contienen

conceptos léxicos y potencialmente pueden tener una definición asociada que puede ser identificada a partir de la consideración de patrones léxico-sintácticos.

1.6.2 Extracción de información

La revolución digital y la disposición actual de enormes repositorios de texto, accesibles prácticamente para todo el mundo, nos ha conducido a un escenario que se torna cada vez más complejo debido al ritmo de crecimiento exponencial que tiene, por ejemplo, la Web. Afortunadamente, se han desarrollado áreas que han llegado a ser muy importantes en la actualidad y que se enfocan en la recuperación y extracción de información a partir de fuentes textuales. En el caso concreto del área de extracción de información (EI), una subdisciplina de la inteligencia artificial, el objetivo es estructurar y combinar datos que se declaran o implican explícitamente en fuentes textuales no estructuradas y/o semi-estructuradas. De acuerdo con Riloff y Lorenzen (1999), un sistema de EI extrae información de textos en lengua natural y para un dominio específico, donde se debe definir previamente el dominio y los tipos de información de interés.

El área de EI se ha asociado tradicionalmente con la extracción de información de fuentes textuales para el llenado de plantillas sobre eventos específicos. Esta tarea se popularizó en las conferencias MUC (Message Understanding Conferences) a finales de los 80's. Uno de los primeros enfoques consistía en técnicas de correspondencia de patrones para llenar plantillas con las instancias de eventos. Actualmente, los enfoques se han enriquecido para incluir técnicas de aprendizaje máquina y estadísticas.

Datos no estructurados

Es común encontrar en la literatura que la tarea de EI se enfoca principalmente en extraer información de fuentes textuales no estructuradas. El modificador *no estructurado* no significa que los datos sean incoherentes estructuralmente, lo que ocurre es que las computadoras no pueden interpretar la información debido a que no se encuentra codificada y dispuesta para su interpretación automática. Es justo en el escenario anterior donde el proceso de EI adquiere relevancia proporcionando significado a datos no

estructurados. Como resultado de un proceso de EI, los datos se convierten a un formato (semi-)estructurado y pueden procesarse de forma fácil y eficaz por una computadora.

En este trabajo entendemos las fuentes de datos no estructurados como textos de lengua natural escrita. Los textos son básicamente de un dominio específico, por ejemplo, medicina, biología, lingüística, etc., y pueden ser de géneros diferentes: diccionarios, enciclopedias, libros, noticias relacionadas con la temática del dominio especializado, etc. Además, asumimos que los textos están bien formados, es decir, que son en buena parte coherentes y libres de error. Lo anterior puede distar mucho de la realidad, sin embargo, creemos que en textos especializados los casos de información incoherente, errores ortográficos y gramaticales son menos comunes que en otros contextos de lengua general.

Extracción de información específica

Como Riloff y Lorenzen lo señalan, EI se aplica en casos donde se conoce previamente qué tipo de información semántica se desea extraer de textos. Por ejemplo, para la extracción de relaciones léxico-semánticas generalmente se determina qué tipo de relación es de interés y de acuerdo con esta selección se construyen manualmente o *aprenden* los patrones lingüísticos que caracterizan la relación y, posteriormente, se recuperan de la fuente textual las instancias.

Uno de los ejemplos más populares de EI es el reconocimiento de entidades nombradas. Las entidades nombradas son nombres de personas, organizaciones, ubicaciones, fechas, dinero, porcentajes, marcas, nombres de proteínas, etc. Por otro lado, otras tareas de EI importantes son la extracción de eventos, donde el énfasis está en obtener los participantes y la configuración de un evento. Del mismo modo, la extracción de escenarios consiste en vincular eventos individuales en una línea histórica. Finalmente, pero no por ello menos importante, está la resolución de correferencia que consiste en determinar si dos expresiones en lenguaje natural refieren a la misma entidad, persona, tiempo, lugar y evento en el mundo. Las tareas anteriores tienen un impacto importante en la investigación sobre EI y siguen definiendo sus objetivos principales.

1.6.3 Información conceptual

Si asumimos que las palabras denotan conceptos, entonces podemos encontrarlos en fuentes de información textual. Actualmente, la lexicografía computacional y la terminología son capaces de reconocer conceptos en corpus de tamaño considerable. Un punto importante que debe considerarse es la fuente de información más apropiada para realizar la extracción de conceptos. En este sentido, Sager (1990) y Smith (2004) señalan el valor de la información científica y técnica como fuentes apropiadas para realizar esta tarea. En particular, Sager (1990) considera las definiciones como una representación lingüística de conceptos debido a que sintetizan toda la información conceptual relacionada con un término en el marco de un dominio de conocimiento específico. Esta postura concuerda con la de Barsalou (2003), porque las definiciones (en este caso particular, definiciones especializadas) trascienden el nivel particular de nuestras experiencias cotidianas.

Por otro lado, Riloff y Shepherd (2004), así como Buitelaar *et al.* (2005), argumentan sobre las dificultades de encontrar información conceptual en corpus de lengua general. Una solución práctica consiste en explorar corpus específicos de dominio, porque contienen información conceptual y léxica que pertenece a un tema concreto.

En congruencia con los autores anteriores, en esta tesis se considera como información conceptual aquella codificada mediante definiciones analíticas de nombres, particularmente definiciones analíticas constituidas por un término, el que se define, un *genus* y una o más *differentiae*, de acuerdo con Smith (2004) y Wilks *et al.* (1996). Estos autores han usado este tipo de definición para buscar relaciones de hiponimia-hiperonimia entre términos y *genus*.

Para reconocer estas relaciones, en Wilks *et al.* (1996) utilizan el operador IS-A para encontrar patrones léxico-sintácticos con un alto grado de precisión, por ejemplo: *Cuchillo: Un borde cortante con un mango que se usa para cortar…*, particularmente en un corpus generado a partir de un diccionario electrónico (ing. Machine-Readable Dictionary, o MRD). Sin embargo, con base en el trabajo realizado por Sierra *et al.* (2008), se deriva que este tipo de patrones no son suficientes para describir todas las posibilidades

de expresar una definición analítica en lenguaje natural. Así, es necesario considerar otros patrones alternativos que sean capaces de introducir estas definiciones en textos especializados.

1.6.4 Extracción de información conceptual

En Sierra *et al.* (2008) se desarrolló un método basado en patrones para extraer términos y definiciones en español. Estas definiciones se encuentran expresadas en fragmentos textuales insertos en documentos especializados. Estos fragmentos se denominan contextos definitorios (CDs), y sus constituyentes principales son un término, una definición y formas lingüísticas o paralingüísticas, tales como frases verbales, patrones pragmáticos y/o marcadores tipográficos, por ejemplo:

> La **energía primaria**, en términos generales, se define como aquel recurso energético que no ha sufrido transformación alguna, con excepción de su extracción.

En el ejemplo anterior se puede ver un CD formado por el término *energía primaria*, la definición *aquel recurso*... y el patrón verbal *se define como*, así como también otras unidades características tales como un patrón pragmático: *en términos generales* y el marcador tipográfico (fuente en negritas) que en este caso enfatiza la presencia del término.

Para lograr el objetivo de la extracción de CDs se emplean patrones verbales que operan como conectores entre términos y definiciones. Dichos patrones sintácticamente son frases predicativas, configuradas alrededor de un verbo que opera como núcleo de esta frase predicativa. Entre los verbos que trabajan como núcleos de frases predicativas, el verbo *ser* es el usado con más frecuencia, principalmente porque permite estructurar operadores tales como el IS-A. Sin embargo, otros verbos pueden fungir como núcleos de estas frases predicativas, es el caso del verbo *definir*, *denominar*, *conocer*, *referir*, etc. La tabla 5 muestra el conjunto de verbos propuestos como frecuentes en definiciones analíticas (Sierra *et al.*, 2008) y que son considerados también en este trabajo. En este contexto, el elemento ε significa la cadena vacía. Con la finalidad de ejemplificar este paradigma analítico donde coexisten verbos como núcleos de frase predicativa, términos y definiciones, se muestran los siguientes casos:

1. La *conjuntivitis* es una *inflamación* de la conjuntiva del ojo.
2. Se define *conjuntivitis* como una *inflamación* de la conjuntiva del ojo.
3. Se denomina *conjuntivitis* a una *inflamación* de la conjuntiva del ojo.

Tabla 5. Predicaciones verbales y elementos relacionados.

Verbo en infinitivo	**Elemento relacionado**
Ser	
Caracterizar, concebir, considerar, describir, definir, entender, conocer, referir	Como
Denominar, llamar, nombrar	Como, ε

En los tres ejemplos anteriores se observa el término y las definiciones relacionadas a través de los núcleos de frase predicativa *ser, definir, denominar*. En todos los casos, el término *conjuntivitis* se concibe como una *inflamación de la conjuntiva*, por esta razón, el *genus* de esta definición es *inflamación*. De acuerdo con Wilks *et al.* (1996), estos casos son un ejemplo canónico de relaciones de hiponimia-hiperonimia en definiciones analíticas.

Por último, resulta importante mencionar que Alarcón (2009) desarrolló un sistema automático denominado ECODE (Extractor Automático de Contextos Definitorios), que reconoce y extrae CDs en español a partir de textos especializados. ECODE identifica y extrae varios tipos de CDs: analíticos, sinonímicos, funcionales[9] y extensionales (Sierra *et al.*, 2008).

1.7 Estructura de la tesis

La organización de esta tesis inicia, en primera instancia, con una descripción del punto de partida y la motivación que da pie a este trabajo de investigación. Posteriormente, se presenta una revisión del estado del arte y, a partir de ésta, se especifican los objetivos, las hipótesis de investigación y las aportaciones de este trabajo al estado del arte actual. Aunado a lo anterior, se considera pertinente proporcionar un resumen de las decisiones más importantes

[9] En la estructura Qualia, Pustejovsky denomina este tipo de información como rol télico.

asumidas desde el inicio de este trabajo.

El capítulo 2 proporciona la base para entender los retos que se enfrentan en la extracción de información conceptual y de relaciones paradigmáticas como la hiponimia-hiperonimia, a partir de textos, incluso para un dominio específico.

En primer lugar, las secciones 2.1 y 2.2 dan cuenta, desde un enfoque clásico, de lo que son los conceptos y las funciones que éstos tienen en nuestros procesos cognitivos, así como también de la conveniencia de esta visión para la formalización de conocimiento. Para complementar la visión de conceptos y porque están estrechamente vinculados, la sección 2.3 explica lo que son las categorías, también desde un enfoque clásico.

Como resultado de las fuertes críticas al modelo clásico, se han propuesto teorías alternativas. La sección 2.4 proporciona una explicación de la teoría de prototipos que ha sustituido, por lo menos en psicología cognitiva, la visión clásica de conceptos. La sección 2.5 proporciona una cronología de la transición mecánica a la computacional en la representación de conceptos mediante primitivos, que se relaciona directamente con una visión clásica y permite vislumbrar la transición de la visión clásica a una basada en prototipos, donde todavía se tiene un terreno fértil y ávido de resultados para implementar computacionalmente. Definir el concepto de *concepto* sigue siendo un tema de enorme debate, sin embargo, resulta necesario adoptar una visión específica cuando éstos se ven involucrados. La sección 2.6 explica cuál es la postura de los que construyen ontologías como recursos computacionales y las críticas que dan lugar a una visión más formal de ontología. Finalmente, concluimos el capítulo con un resumen sobre los aspectos más relevantes de los contenidos presentados.

El capítulo 3 presenta un panorama general de lo que son las relaciones léxico-semánticas desde la perspectiva de la semántica léxica. La sección 3.1 da una explicación general de lo que son las relaciones sintagmáticas y paradigmáticas y cómo contribuyen al significado de las palabras. La sección 3.2 explica el concepto de *sentido de palabras* como un objeto semántico influenciado por la ambigüedad del lenguaje y la imposibilidad de crear categorías con límites claros y precisos. Por otro lado, la sección 3.3 explica la

insuficiencia de las relaciones paradigmáticas para cubrir toda la amplia gama de asociaciones entre conceptos. Por su parte, la sección 3.4 describe las relaciones paradigmáticas más exploradas en la literatura, en términos lógicos y lingüísticos para el caso de las relaciones de hiponimia-hiperonimia y meronimia-holonimia, así como su implementación para organizar la base de datos léxica WordNet. Finalmente, La sección 3.5 proporciona una revisión general de los formalismos de representación de conocimiento usados en inteligencia artificial que consideraron los beneficios de las jerarquías conceptuales.

El capítulo 4 corresponde a una descripción detallada de la metodología propuesta en nuestra investigación. En este apartado se discuten los procesos llevados a cabo en el corpus, así como también las heurísticas aplicadas para identificar y filtrar la información relevante. Posteriormente, el capítulo 5 enfatiza y discute los resultados encontrados con la aplicación de la metodología a una fuente textual específica.

Para concluir, el capítulo 6 prioriza las conclusiones derivadas de los experimentos, así como también del trabajo futuro que complementaría de forma eficaz la metodología propuesta. Al final, se resaltan las contribuciones principales de esta investigación al estado del arte actual.

"(...) Esas ambigüedades, redundancias y deficiencias recuerdan las que el doctor Franz Kuhn atribuye a cierta enciclopedia china que se titula Emporio celestial de conocimientos benévolos. En sus remotas páginas está escrito que los animales se dividen en (a) pertenecientes al Emperador, (b) embalsamados, (c) lechones, (e) sirenas, (f) fabulosos, (g) perros sueltos, (h) incluidos en esta clasificación, (i) que se agitan como locos, (j) innumerables, (k) dibujados con un pincel finísimo de pelo de camello, (l) etcétera, (m) que acaban de romper el jarrón, (n) que de lejos parecen moscas."

Jorge Luis Borges, "El idioma analítico de John Wilkins", en Obras Completas, Vol. II (1952-1972), Barcelona, Emecé Editores, 1996, p. 85-86.[10]

Rosch (1978) opina sobre la taxonomía del reino animal descrita anteriormente:

"(…) Conceptually, the most interesting aspect of this classification system is that it does not exist. Certain types of categorizations may appear in the imagination of poets, but they are never found in the practical or linguistic classes of organisms or of man-made objects used by any of the cultures of the world."

"(…) Well, I'll tell you something. You really don't know what a metal is. And there's a big group of people that don't know what a metal is. Do you know what we call them? Metallurgists! . . . Here's why metallurgists don't know what metal is. We know that a metal is an element that has metallic properties. So we start to enumerate all these properties: electrical conductivity, thermal conductivity, ductility, malleability, strength, high density. Then you say, how many of these properties does an element have to have to classify as a metal? And do you know what? We can't get the metallurgists to agree. Some say three properties; some say five properties, six properties. We really don't know. So we just proceed along presuming that we are all talking about the same thing."

The big book of concepts, Gregory L. Murphy (2002)

[10] Publicado originalmente en: Jorge Luis Borges (1952): *Otras Inquisiciones*, Sur, Buenos Aires, pp. 85-86.

Capítulo 2

Conceptos y categorías

La definición de lo que es un *concepto* sigue siendo un tema de fuertes debates. Por ejemplo, en disciplinas como la filosofía y la psicología cognitiva no existe un consenso aún de lo que son, cómo se forman o si realmente existen; sin embargo, se han propuesto teorías que se enfocan en explicar estos tres cuestionamientos principales. Una de las teorías que mayor influencia ha tenido es la teoría clásica (Aristotélica) que plantea la construcción de la definición de un concepto a partir de determinar propiedades necesarias y suficientes. Sin embargo, esta postura ha generado fuertes críticas tanto en el ámbito filosófico, como en el de la psicología cognitiva (Smith y Medin, 1981; Croft y Cruse, 2004; Murphy, 2002; Lakoff, 1987).

Dado que en este trabajo partimos de una etapa de extracción conceptual, concretamente de la extracción de CDs analíticos localizados en textos especializados, consideramos importante presentar una exploración general de lo que son los conceptos y de un elemento que está estrechamente relacionado con ellos y, de manera directa, con definiciones analíticas: la categorización. Adoptamos una visión cognitiva de conceptos y categorías dada la influencia que ha tenido la psicología cognitiva en el enfoque de adquisición de conocimiento en corpus (Poesio, 2005). De forma breve y clara, intentamos explicar estos dos conceptos desde una perspectiva clásica, dando cuenta de las ventajas, así como también de las desventajas que se han evidenciado de su aplicación a la realidad. Posteriormente, presentamos una descripción general del sistema conceptual jerarquizado propuesto por Rosch (1978), visto desde dos perspectivas: vertical y horizontal. Como investigaciones derivadas del enfoque anterior, describimos resultados de estudios de la psicología cognitiva que consideran el *genus* como el nivel básico, así como también otros que indican que el nivel de conocimiento de un dominio tiene relación directa con el nivel seleccionado: a grandes rasgos, mientras más conocimiento o experiencia se tenga en un dominio, mayor tendencia a seleccionar categorías subordinadas o más específicas.

Por otro lado, resulta de gran interés presentar una cronología de transición del enfoque mecánico al computacional de representación de conceptos mediante primitivos, que nos permita llevar un seguimiento general de la necesidad de migrar de un enfoque clásico a uno basado en prototipos (Sowa, 2005).

Por último, presentamos la visión de conceptos en el marco de las ontologías como recursos computacionales y las críticas respecto a esta postura. En un intento por complementar lo anterior, concluimos con la presentación de una metodología para construir ontologías de fuentes textuales.

2.1 Conceptos

De acuerdo con Smith (1988), la noción de concepto es esencial para comprender el pensamiento y el comportamiento humano. Los conceptos reflejan la forma en que dividimos el mundo en clases, y mucho de lo que aprendemos, comunicamos y razonamos incluye relaciones entre estas clases. En términos de la psicología cognitiva, un concepto se considera como una representación mental de una clase que nos ayuda a ver las cosas o experiencias nuevas como similares a lo que ya conocemos (Murphy, 2002). Esta habilidad conceptual de relacionar el presente con el pasado nos evita manipular enormes lexicones mentales que, de ser el caso, harían prácticamente imposible nuestros procesos de razonamiento y comunicación debido a que tendríamos que reconocer y nombrar cada elemento o experiencia como algo diferente. Dada la importancia que tienen los conceptos, resulta imprescindible hacer mención de sus funciones más relevantes:

1. Promover la economía cognitiva mediante la clasificación de todo lo que existe en el mundo, de tal suerte que podamos hacer más eficientes nuestros procesos de razonamiento y comunicación.
2. Habilitar la gestión de mayor información que la proporcionada en un determinado momento. Por ejemplo, cuando escuchamos la palabra "silla"[11] (que denota el concepto *silla*), inmediatamente podemos inferir más propiedades sobre el concepto: tiene 4 patas,

[11] Representamos las palabras entre comillas y el concepto que denotan mediante cursivas.

sirve para sentarse, puede ser de madera o metal, etc.

3. Combinar conceptos simples para formar conceptos y pensamientos complejos. Por ejemplo, el concepto *sombrero rojo* involucra aquellos elementos que pertenecen a la clase *sombrero* porque se usan como prenda de vestir y pertenecen también a la categoría de cosas que son *rojas*.

2.2 Teoría clásica de conceptos

De acuerdo con Sowa (2005), Aristóteles propuso dos métodos de clasificación: un método de arriba hacia abajo (top-down), que consiste en definir conceptos basados en un *genus* o supertipo y una o más *differentiae*. Por otro lado, formuló otro método, de abajo hacia arriba (bottom-up), que inicia con una descripción detallada del objeto o entidad, clasificándolo gradualmente por especies y estas especies por géneros.

La teoría clásica sostiene que la mayoría de los conceptos, especialmente los lexicalizados, tienen una estructura definicional, es decir, se codifican en términos de condiciones necesarias y suficientes (Smith y Medin, 1981; Laurence y Margolis, 1999), lo que se relaciona directamente con el primer enfoque top-down mencionado en el primer párrafo. Por ejemplo, el concepto *soltero* se puede pensar como una representación mental compleja que especifica condiciones necesarias y suficientes para que algo sea clasificado como un *soltero*. Así, el concepto *soltero* podría estar conformado por un conjunto de rasgos o primitivos tales como: *libre*, *hombre*, *adulto*. Cada uno de estos rasgos especifica una condición que cualquier entidad debe cumplir para ser un *soltero*, de manera que cualquier cosa o entidad que las satisfaga, por tanto, debe ser un *soltero*. Estos rasgos producen una interpretación semántica para la representación de conceptos complejos acorde con los principios de la semántica composicional.

Una de las ventajas más reconocidas del enfoque clásico está en su potencial de formalización de conocimiento. Con el objetivo de representar la conveniencia de este enfoque clásico para representación de conocimiento y su razonamiento, consideramos la explicación que proporciona Sowa (2005) del comentario que el filósofo Porfirio (siglo III D.C) realizó sobre las categorías de Aristóteles. Para facilitar la explicación, se extrajo la figura 4 de Sowa (2005).

Esta imagen representa las categorías y sus relaciones con silogismos que Aristóteles usaba para razonar respecto a tipos y subtipos. A partir del esquema queda claro que con el *genus* supremo Substance, la *differentia* Material es el subtipo Body, y con la *differentia* Inmaterial es Spirit. En este sentido, la herencia es el proceso de unir la *differentia* siguiendo una ruta específica. Por ejemplo, LivingThing se define como "animate material Substance", y Human es "rational sensitive animate material Sustance".

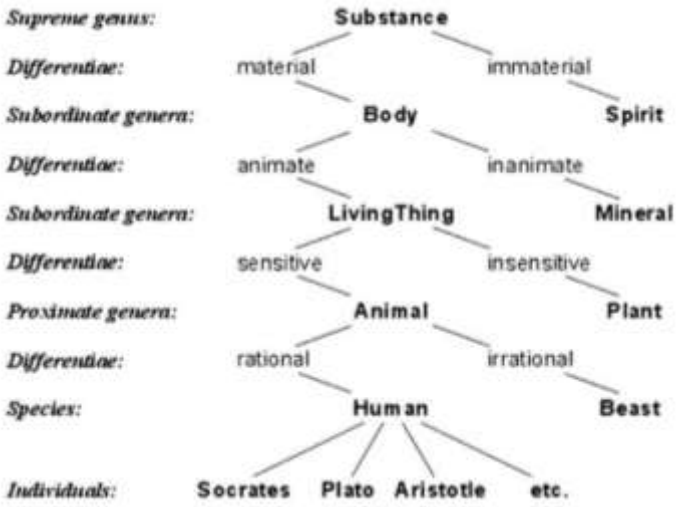

Figura 4. Árbol de Porfirio, extraído de Sowa (2005).

Por otro lado, la desventaja principal es la incapacidad de enunciar una definición en términos de condiciones necesarias y suficientes para una gran cantidad de conceptos. El filósofo Wittgenstein (1953) cuestionó la visión clásica para construir definiciones y propuso como ejemplo el concepto *juego*. Es realmente difícil definir el concepto *juego* de tal suerte que todos los que pertenecen a este concepto queden incluidos y que además se excluyan aquellas cosas que no lo son. La estrategia de Wittgenstein fue pedir a sus lectores que no sólo pensaran: "debe haber algo en común entre todos los miembros", sino que trataran de especificar las características comunes. De esta forma dio cuenta de que puede llegar a ser muy difícil especificar rasgos necesarios y suficientes para muchos conceptos.

Cabe mencionar aquí que el argumento de Wittgenstein se acepta ampliamente en el campo de la psicología cognitiva; sin embargo, de acuerdo con Smith y Medin (1981), no es del todo verdadero. Estos autores argumentan que cuando decimos que no podemos pensar en rasgos que definan el concepto

juego, esto no prueba que no existan, simplemente evidencia que no somos lo suficientemente capaces de definirlos. Lo anterior puede ser, entre otras cosas, porque estas propiedades no se han descubierto aún, lo que habla de la evolución del conocimiento respecto a las cosas en el mundo (Sowa, 2005).

2.3 Categorías

La categorización es una de las actividades cognitivas humanas más básicas e importantes. De acuerdo con Smith y Medin (1981), en psicología, las investigaciones sobre categorización son las que han promovido la transición de un enfoque clásico a uno probabilístico. El proceso de categorización incluye la aprehensión de alguna entidad nueva como parte de algo que ya se ha concebido de forma abstracta y que incluye también otras instancias reales o potenciales (Croft y Cruse, 2004).

Conceptos y categorías son dos elementos que no pueden verse de forma independiente. Para Smith y Medin (1981), decir que los conceptos tienen una función de categorización es reconocer que son mecanismos de reconocimiento de patrones, lo que significa que los conceptos se usan para clasificar entidades nuevas y extraer inferencias respecto a ellas. Si tengo un concepto de X, entonces sé algo relacionado con las propiedades de las entidades que pertenecen a la clase X, y estas propiedades las puedo usar para categorizar objetos o entidades nuevas. Por otro lado, si no sé nada respecto a un objeto nuevo pero alguien me dice que es una instancia de X, puedo inferir que el objeto tiene todas o muchas de las propiedades de X, ésto implica ver el proceso de categorización de forma inversa. Entre las funciones básicas de las categorías se encuentran:

- Aprendizaje. Las experiencias nunca ocurren exactamente igual, por ello resulta indispensable aprender de las experiencias pasadas para aplicar aspectos similares a las experiencias nuevas y poder colocarlas en las mismas categorías conceptuales.
- Planificación. La formulación de metas y planes para lograr estas metas requiere disociar conocimiento de individuales y empaquetarlos en conceptos que caractericen categorías de entidades.
- Comunicación. El lenguaje trabaja en términos de categorías. Cualquier

expresión, aunque sea detallada, al final representa solo una categoría de referentes.

- Economía. El conocimiento con frecuencia no se relaciona con individuales, en lugar de ello, se relaciona con grupos de individuales. Así, es posible obtener conocimiento nuevo mediante la interacción con uno o más individuales y generalizarlo a los otros miembros de la categoría. Del mismo modo, saber que un individuo pertenece a una categoría específica nos da acceso a mayor información respecto al individual.

2.3.1 El modelo clásico de categorías

El enfoque clásico establece un límite bien definido y rígido para un concepto. Para ilustrar la visión clásica consideremos el ejemplo propuesto por Smith y Medin (1981) del concepto geométrico de un *cuadrado*. Supóngase que las personas, en general, representan el concepto cuadrado en términos de 4 propiedades básicas: 1) figura cerrada, 2) tiene cuatro lados, 3) los lados son de igual longitud y 4) los ángulos son iguales. Bajo el enfoque clásico, para determinar el estatus de cuadrado de cualquier figura se validarían las 4 propiedades anteriores, por tanto, podemos decir que tenemos una descripción unitaria del concepto *cuadrado*. En términos generales, una descripción unitaria es el conjunto de propiedades necesarias y suficientes que todos los miembros de la clase deben cumplir. Además, cualquier objeto o entidad en el mundo que las cumpla debe ser un miembro de la clase.

En la perspectiva clásica, las relaciones de inclusión entre categorías se encuentran bien reconocidas, sin embargo, no existe explicación sobre niveles absolutos de categorización. De acuerdo con Croft y Cruse (2004), el modelo clásico presenta varios problemas. Tres de los más citados en la literatura y que han estimulado el desarrollo de teorías alternativas, son los siguientes:

- La dificultad de establecer una serie de condiciones necesarias y suficientes para conformar una definición adecuada de un concepto, el ejemplo más famoso es aquel proporcionado por Wittgenstein del concepto *Juego*[12]. Aunado a lo anterior, incluso para los conceptos que

[12] Para más detalle, revisar: http://plato.stanford.edu/entries/wittgenstein/#Lan

parecen tener una definición, ésta con frecuencia es válida sólo dentro de un dominio específico (Fillmore, 1975).

- Algunos miembros de una categoría son más representativos de la categoría que otros (Rosch, 1978). En el modelo clásico, se asume que todos los miembros son iguales.
- El modelo clásico no ofrece ninguna explicación del por qué, en la práctica, los límites de las categorías parecen ser difusos y variables.

Como se mencionó en párrafos anteriores, la teoría clásica representa una perspectiva para la formación de conceptos que ha sido fuertemente cuestionada. A raíz de los cuestionamientos mencionados anteriormente, se han propuesto teorías alternativas. Una de estas teorías es la de prototipos. La siguiente sección muestra un panorama general de este enfoque de prototipos.

2.4 La teoría de prototipos

La teoría de prototipos asume que las instancias de un concepto varían en el grado en el que comparten ciertas propiedades, y en consecuencia varían en el grado en el que representan el concepto. Con la finalidad de ilustrar esta postura, consideramos el ejemplo propuesto por Smith y Medin (1981) del concepto *taza,* tratando de conceptualizarlo desde una perspectiva clásica. Las propiedades que podrían formar parte de la descripción unitaria de este concepto son: 1) objeto concreto, 2) cóncavo, 3) puede contener líquidos, 4) tiene una manija, y 5) sirve para tomar líquidos. La pregunta que surge a partir de esta descripción es: ¿todas las propiedades son verdaderas para todo lo que la gente denomina como *taza*? Las propiedades de la 1 a la 3 parecen ser los rasgos más preponderantes en *tazas*, pero no 4 y 5, que podrían ser verdaderas solo para un subconjunto de estos objetos. Sin embargo, si dejamos de lado las dos últimas propiedades y nos quedamos con las tres primeras, éstas aplican también a objetos que no son tazas, por ejemplo, *tazones*. De lo anterior se deduce el hecho de que, para una gran cantidad de conceptos, las propiedades que se formulen en una descripción unitaria no siempre serán aplicables a todos los miembros de una clase, por lo que es necesario postular una descripción donde las propiedades sean verdaderas para una buena parte de los miembros de una clase. El análisis de casos como

el anterior fue precisamente lo que dio origen a la teoría de prototipos o, como Smith y Medin (1981) asumen denominarla, la visión probabilística de conceptos.

En las siguientes secciones se presenta una revisión general de la perspectiva de Rosch (1978) de lo que subyace a los sistemas de categorización humana. En primera instancia, se describen los dos principios básicos implícitos en los sistemas de categorización y cómo aplican en las dos dimensiones de un sistema categorial: estructura jerárquica (dimensión vertical) y estructura de categorías dentro de un nivel específico (dimensión horizontal). Por otro lado, con el objetivo de ofrecer una mayor claridad respecto a las dos dimensiones mencionadas anteriormente, se incluyen dos apartados con una explicación breve.

Finalmente, derivado del trabajo de Rosch y Mervis (1975), se realizó una gran cantidad de investigación para obtener evidencia empírica y con ello soportar las críticas contra el modelo clásico. Prueba de lo anterior son dos resultados incluidos en un último apartado que dan cuenta sobre la relación entre la categoría de nivel básico con el *genus* y categorías subordinadas.

2.4.1 Principios de categorización

El estudio del sistema conceptual humano se originó a partir de los trabajos llevados a cabo por Rosch y colegas en los 70s. La existencia de un sistema jerárquico de categorías despertó gran interés. Rosch (1978), en su artículo Principios de categorización (*Principles of Categorization*), propuso dos principios básicos y generales para la formación de categorías.

El primer principio se refiere a la función de los sistemas categoriales y afirma que la tarea de estos sistemas es proveer de máxima información con el menor esfuerzo cognitivo. El segundo principio aborda la estructura de la información y afirma que el mundo percibido no es un conjunto de atributos impredecibles y no correlacionados, por el contrario, tiene estructura. Así, la información máxima con el menor esfuerzo cognitivo se logra si las categorías reflejan la estructura del mundo percibido lo mejor posible.

De acuerdo con Rosch, el principio de economía cognitiva combinado con la estructura del mundo percibido tiene implicaciones importantes en el nivel de abstracción de las categorías formadas por una cultura y en la

estructura interna de esas categorías una vez que se forman.

Para desglosar mejor su explicación respecto a lo anterior, plantea ver el sistema conceptual desde dos perspectivas, una vertical y otra horizontal. La dimensión vertical se refiere al nivel de inclusividad de la categoría y básicamente es la relación de subsunción entre diferentes categorías, por ejemplo:

$$\text{Poodle} \subset \text{perro} \subset \text{mamífero} \subset \text{animal}$$

Las implicaciones de los dos principios de categorización para la dimensión vertical consiste en que no todos los niveles de categorización son igualmente útiles, es decir, existe un nivel de categorización básico, que es el nivel más inclusivo en el que las categorías pueden reflejar la estructura de los atributos percibidos en el mundo. Este nivel de inclusividad resultó ser la parte media entre los niveles más inclusivos y los menos inclusivos, es decir, el nivel asociado con categorías como *carro*, *perro* y *asiento* (ver figura 5).

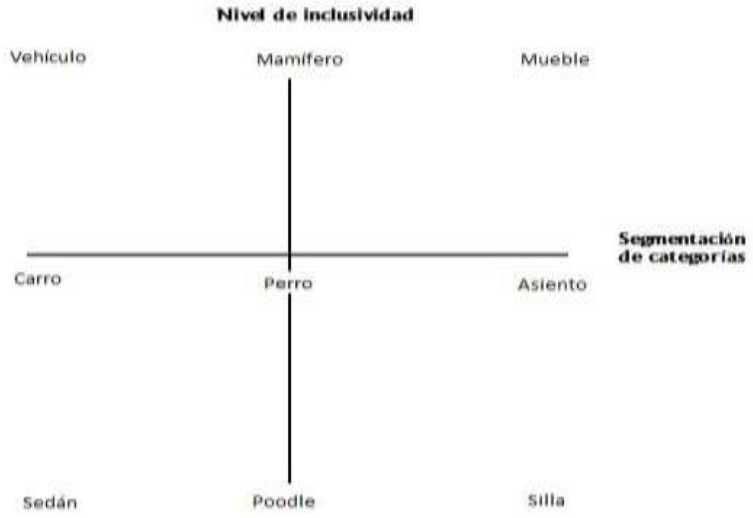

Figura 5. El sistema de categorización humana.
Adaptada de Evans y Green (2006).

Por otro lado, la dimensión horizontal hace énfasis en la segmentación de categorías en el mismo nivel de inclusividad, es decir, la dimensión en la que *perro*, *carro* y *asiento* varían. Las implicaciones de los principios de categorización para la dimensión horizontal consisten en que al incrementar el nivel de distinción y la flexibilidad de las categorías, éstas tienden a ser definidas en términos de prototipos, mismos que contienen los atributos más

representativos de los elementos dentro de la categoría, y menos representativos de los elementos de otras categorías. Esta dimensión horizontal se relaciona, en particular, con el principio de estructura del mundo percibido. Como se mencionó anteriormente, el principio afirma que el mundo tiene estructura y esta estructura proporciona restricciones en los tipos de categorías que los humanos representan dentro de su sistema cognitivo.

2.4.2 La dimensión vertical

Las categorías se encuentran en diferentes niveles de inclusividad, con las más específicas anidadas dentro de las más inclusivas. Los siguientes son algunos ejemplos extraídos, traducidos y adaptados de Croft y Cruse (2004):

1. Sedán $\subset$ carro $\subset$ vehículo.
2. Granny Smith $\subset$ manzana $\subset$ fruta.
3. Spaniel $\subset$ perro $\subset$ animal $\subset$ criatura $\subset$ cosa viviente.

De acuerdo con Rosch *et al.* (1976), el nivel básico es el que tiene mayor importancia y un estatus especial. Existen dos niveles más que están supeditados al nivel básico: el nivel superordinado y el subordinado. Por ejemplo, si el nivel básico del ejemplo 3) es la categoría *perro*, el nivel superordinado será *animal* y el subordinado será *Spaniel*.

La siguiente descripción de categorías, en cada uno de los niveles de un sistema conceptual jerarquizado, se realiza con base en Croft y Cruse (2004).

Categoría de nivel básico

En los experimentos realizados por Rosch y colaboradores en los 70s se encontró que un nivel de abstracción básico tenía un estatus especial en el proceso de categorización humana. El nivel básico mostró ser el más inclusivo en el que es posible identificar e imaginar formas generales de los ejemplares de una categoría. Las siguientes características corresponden al nivel básico:

- Es el nivel más inclusivo en el que existen patrones de comportamiento característicos.
- Es el nivel más inclusivo para el que se puede formar una imagen visual clara.
- Es el nivel más inclusivo en el que se puede representar información parte-todo.

- Es el nivel que se utiliza como referencia neutral.
- Los elementos individuales se categorizan más rápidamente como miembros de categorías básicas que como miembros de categorías superordinadas o subordinadas.

Categoría de nivel superordinado

Las características más importantes de las categorías superordinadas son las siguientes:

- Agrupan elementos que son relativamente diferentes de otras categorías vecinas, sin embargo, incluso la semejanza dentro de una categoría específica es baja. Por ejemplo, las categorías *planta* y *animal* agrupan entidades diferentes. De manera semejante, dentro de la clase *animal*, si comparamos un *pez* con un *perro* o un *ave*, estos tienen poca semejanza.
- Tienen menos atributos que definan la categoría comparada con las categorías de nivel básico. Por ejemplo, la categoría *perro* tiene más rasgos característicos que la categoría *animal*.

Categorías de nivel subordinado

Las características de las categorías de nivel subordinado son:

- Menos útiles que las de nivel básico, porque aunque los miembros tienen un alto parecido, se distinguen menos de los miembros de categorías vecinas.
- Menos informativas comparadas con su categoría hiperonímica inmediata, por tanto, cuando se les pide a las personas listar atributos distintivos, las listas difieren muy poco de las listas del nivel básico hiperonímico.

2.4.3 Dimensión horizontal (centralidad gradual)

Los miembros de una categoría difieren en cuanto a lo representativo que pueden ser de ella. En algunas ocasiones intuimos que algunos miembros son mejores ejemplos que otros. Por tanto, los más representativos estarían en el núcleo de la categoría. En el marco de la teoría de prototipos, los psicólogos cognitivos han realizado mucho trabajo experimental sobre la noción de

bondad del ejemplar (Goodness-Of-Exemplar, o GOE). Entre los estudios más simples se encuentran aquellos que muestran una lista de elementos de una categoría y piden asignar a cada miembro un número de acuerdo a qué tan buen ejemplo es de esa categoría. La tabla 6 muestra varios elementos de la categoría *vegetales,* donde la puntuación menor corresponde al mejor ejemplo. Así, los mejores ejemplos en este caso son *puerro* y *zanahoria*. Los resultados de estos experimentos, cuando se aplican a un gran número de sujetos, permiten la identificación de los mejores ejemplos de categorías (*prototipos* o *miembros prototípicos).*

Tabla 6. Puntajes GOE de miembros de la categoría vegetales (Croft y Cruse, 2004)

Vegetales	Puntaje GOE
Puerro, zanahoria	1
Brócoli, chirivía	2
Apio, remolacha	3
Berenjena, calabacín	4
Perejil, albahaca	5
Ruibarbo	6
Limón	7

La significancia de los puntajes GOE se ha ampliado considerando los experimentos que muestran su correlación significativa con propiedades tales como la frecuencia y orden de mención de un elemento, el orden de aprendizaje, parecido de familia (family resemblance), velocidad de verificación y experimentos *priming*[13]. La velocidad de verificación y *priming* se han considerado correlacionados significativamente con puntuaciones GOE porque no están bajo el control consciente y por tanto se puede afirmar que revelan propiedades que subyacen a las categorías.

En este tipo de estudios existe otro indicador denominado grado de membresía (en inglés, DOM) en la categoría. Algunos afirman que representan la misma información, lo que es incorrecto. GOE rankea los elementos para ver cuán buenos son como miembros de una categoría específica. Por ejemplo, una *avestruz* es un miembro de la categoría *pájaro*; sin embargo, es innegable que tiene un GOE bajo, por lo que, en este caso, ambos indicadores difieren. Dado

[13] En psicología, *priming* es un efecto relacionado con la memoria implícita por el cual la exposición a determinados estímulos influye en la respuesta que se da a estímulos presentados con posterioridad. Por ejemplo, la palabra *perro* se reconoce más fácilmente si está precedida por la palabra *gato*.

lo anterior, los dos parámetros reflejan aspectos diferentes y por ello deben ser independientes (Véase Murphy, 2002 para una revisión más detallada).

2.4.4 El *genus* como nivel básico de categorización

Murphy (2002) explica la importancia del nivel básico en términos de su nivel de información y distinción. Por un lado, los conceptos básicos se asocian con una gran cantidad de información. Por ejemplo, si se tiene conocimiento de que una entidad es un *perro*, se puede inferir que ladra, tiene 4 patas, tiene pelo, come carne, etc. Por otro lado, su nivel de distinción refiere al hecho de que es diferente de otras categorías en el mismo nivel. Por ejemplo, los *perros* son diferentes de los *gatos*, *caballos*, *vacas* y otros mamíferos.

En psicología cognitiva se han realizado investigaciones sobre el nivel de categorización más usado y los factores que inciden en su uso. Algunos resultados indican que el nivel básico más usado es el *genus*. Uno de los estudios que revelaron este tipo de resultados se hizo con hablantes nativos del lenguaje Tzeltal sobre clasificaciones populares. Los nativos tendían a nombrar plantas y animales en un solo nivel de clasificación científica, el *genus* (pino, róbalo), en lugar de uno más específico (pino blanco, róbalo negro) o uno más general (árbol, pez). Estos resultados sugirieron que la gente, a través de todas las culturas, usaba el *genus* como nivel básico (Berlin, Breedlove y Raven, 1974). Sin embargo, esta propuesta se consideró demasiado rígida. Posteriormente, otros estudios revelaron que el nivel de conocimiento de las personas influye en el nivel considerado. Es decir, una persona que no tiene experiencia ni conocimiento en un dominio tendería a usar categorías superordinadas. Por otro lado, personas con un nivel de entrenamiento y conocimiento más elevado de un dominio usarían niveles subordinados porque, contrario a las características de categorías subordinadas mencionadas en la sección 2.4.2, los expertos conocen rasgos únicos a sus conceptos subordinados, lo que los hace distintivos para ellos (Tanaka y Taylor, 1991).

2.5 La representación de conceptos mediante primitivos

En este apartado presentamos un extracto de la descripción de Sowa (2005) respecto a la transición de enfoques mecánicos hasta las modernas

implementaciones computacionales de representación conceptual mediante primitivos.

Ramon Llull (1303) fue el primero que propuso la idea de relacionar mecánicamente conceptos con primitivos. Llull desarrolló un sistema llamado *Ars Magna* que consistía de un conjunto de discos inscritos con conceptos primitivos, que podían combinarse de varias formas mediante su rotación. Posteriormente, Leibniz (1666) se inspiró en Lull para desarrollar su *Universal Characteristic*, que representaba conceptos primitivos por medio de números primos y utilizaba el producto de estos números primos para construir conceptos más complejos. Leibniz imaginó un diccionario universal para mapear conceptos a números y un motor de razonamiento para automatizar los silogismos.

Con la llegada de las computadoras, los lingüistas computacionales se fijaron como meta implementar el diccionario universal de Leibniz. Por ejemplo, Masterman (1961), definió una red semántica de 15,000 palabras, que organizó como un *lattice* basado en 100 conceptos primitivos, entre ellos estaban: [FOLK], [STUFF], [CHANGE], [GO] y [TALK]. Por ejemplo, para la oración *This man is greedy, but pusillanimous*, el sistema generaba la representación:

(THIS: MAN:) (HE: (CAN/ DO/ (MUCH: EAT)))
(BUT: NOT:) (HE: (CAN/ DO/ (MUCH: FIGHT))).

Por su parte Schank (1975), en su implementación de *grafos de dependencia conceptual*, redujo el número de actos primitivos a ll. Por ejemplo, la frase *x bought y*, se podía expandir a *x obtained posession of y in exchange for money*. Esta transformación de conceptos complejos a primitivos permitía derivar frases sinónimas.

Los diccionarios modernos analizan miles de palabras en primitivos, sin embargo, no se limitan a un conjunto fijo de categorías. La mayoría de los diccionarios tienen definiciones circulares. Por ejemplo, es común encontrar una definición del concepto *atributo* como *característica* y *característica* como *atributo*.

En lingüística, Katz y Fodor (1963) introdujeron primitivos que denominaron *marcadores semánticos* con *reglas de proyección* para

combinarlos. Muchos lingüistas adoptaron variantes de este método, pero incluso los que lo usaron plantearon críticas serias en los sentidos siguientes:

- No había indicios lingüísticos o psicológicos que dieran cuenta de la división brusca entre la información en los marcadores semánticos y la información restante, a la que Katz y Fodor llamaron *distinguisher*.
- La mayoría de los lenguajes contienen familias de sinónimos, cada uno con matices ligeramente diferentes de significado. En su planteamiento, los marcadores semánticos solo soportan dicotomías.
- Los marcadores semánticos, semejante a los números primos de Leibniz, solo pueden representar conjunciones de primitivos. Se requieren otros operadores para representar todas las relaciones lógicas.

Los métodos basados en lógica, que se remontan desde el Árbol de Porfirio hasta los últimos enfoques de ontologías formales, son ejemplos del enfoque Aristotélico top-down. Como muchos detractores del enfoque clásico, Whewell (1858) afirmó que las definiciones top-down en Biología eran inútiles. Por su parte, Mill (1865) desechó el supuesto de condiciones necesarias y suficientes, por un criterio más débil basado en todas las características necesarias y una mayoría opcionales.

Finalmente, Wittgenstein (1953), como se mencionó en secciones anteriores, cuestionó el enfoque clásico al demostrar que palabras ordinarias como "juego" no podían definirse en términos de condiciones necesarias y suficientes. Él propuso que, en lugar de *differentia* que distingue los diferentes tipos de juegos de otros, los juegos comparten un *parecido de familia.*

Los métodos de definición empleados en el siglo XIX se han refinado e implementado en sistemas computacionales modernos:

- Lógicos. Un concepto se define por un *genus* y un conjunto de propiedades necesarias y suficientes. Este método se usa todavía en diccionarios y sistemas de razonamiento deductivo.
- Difuso. Un concepto se define por cero o más condiciones necesarias y una preponderancia de opcionales que se ordenan por importancia. Variaciones del criterio de Mill se han implementado en métodos computacionales basados en redes neuronales así como también en conjuntos difusos, conjuntos aproximativos (rough sets),

análisis de agrupamiento y análisis de semántica latente.

- Prototipo. Un concepto se define por un ejemplo o prototipo, y cualquier instancia del concepto debe parecerse al prototipo del concepto. Este método incluye el *parecido de familia* de Wittgenstein y los prototipos psicológicos de Rosch. Una implementación de computadora requiere alguna medida de distancia entre prototipos, y la búsqueda de medidas apropiadas se ha convertido en un tema de investigación importante.

Para concluir su cronología de transición, Sowa señala que después de dos milenios de debate filosófico y más de 50 años de implementaciones computacionales, el consenso moderno no dista mucho de la postura de Kant (1800):

> Since the synthesis of empirical concepts is not arbitrary but based on experience, and as such can never be complete (for in experience ever new characteristics of the concept can be discovered), empirical concepts cannot be defined. Thus only arbitrarily made concepts can be defined synthetically.
> Such definitions... could also be called *declarations*, since in them one declares one's thoughts or renders account of what one understands by a word. This is the case with *mathematicians*.

2.6 Conceptos en el marco de las ontologías computacionales

En el marco de la filosofía, concretamente de la metafísica, el término ontología designa al estudio que se enfoca en el análisis de las entidades existentes y sus propiedades (Hofweber, 2004). Dicho estudio se enfoca en determinar y explicar de qué modo existen y se comportan los objetos dentro del mundo. Del mismo modo, la ontología también explora los rasgos generales de dichos objetos, junto con las relaciones que éstos establecen entre sí.

Desde un punto de vista más práctico, y considerando la meta de compartir conocimiento, las ontologías desarrolladas desde la perspectiva de las ciencias de la información y la computación han estado centradas en conceptos, entendiendo concepto como un producto de la cognición humana (Smith, 2004).

En consideración con lo expresado en el párrafo anterior, una de las definiciones de ontología más citada en la literatura computacional (Guarino,

1998; Noy y McGuinness, 2001; Smith, 2004; Buiteelar *et al.*, 2005), es la formulada por Gruber (1993). Este autor define ontología como una especificación formal y explícita de una conceptualización compartida respecto a un dominio de interés.

En la definición anterior, se plantea la conceptualización como una visión abstracta y simplificada del mundo, o una parte de éste que deseamos representar para algún propósito. El término *formal* implica que la ontología debe ser procesable por la computadora, y *explícita* se refiere a que el tipo de conceptos usados y sus restricciones deben ser definidos explícitamente. Finalmente, por *compartida* se entiende que debe capturar el conocimiento consensuado y aceptado por los estudiosos o interesados en el dominio que se intenta modelar.

En el mismo orden de ideas, y con la finalidad de mostrar parte de las críticas al enfoque de Gruber o, en general, al enfoque basado en conceptos sobre el desarrollo de ontologías, Smith (2004) plantea que la influencia de esta visión centrada en conceptos es consecuencia no sólo de las raíces de los sistemas de información en el campo de la representación de conocimiento, sino también promovida porque mucho del trabajo sobre ontologías se ha enfocado en la representación de dominios tales como comercio, leyes, o administración pública, donde se trata generalmente con los productos del acuerdo y la convención humana. Smith sostiene la tesis de que lo que se debe modelar no son conceptos como abstracciones o convenciones humanas, sino universales y particulares basados en sólidos principios filosóficos y/o evidencia científica.

La perspectiva de Smith refleja el punto de vista de las ontologías formales contrastado con aquel de las ontologías basadas en conceptos. De acuerdo con Poesio (2005), el origen de esta división se dio a partir de las primeras propuestas de codificación de ontologías mediante redes semánticas y sistemas de marcos (Quillian, 1968; Minsky, 1975). Estos enfoques tuvieron gran aceptación por su facilidad y eficiencia en la implementación de inferencias tales como herencia y similitud. Sin embargo, la mayoría de las implementaciones de estos enfoques de representación se consideraron informales por su falta de precisión semántica y también por el lado ontológico:

¿qué tipo de objetos existen? Estas deficiencias motivaron el desarrollo de una gran cantidad de investigación enfocada a proveer de más rigor a los formalismos de representación. Es justo en este punto donde inician las diferencias de percepción de los problemas, así como también de sus posibles soluciones y se establecen por lo menos las dos tradiciones de investigación mencionadas anteriormente.

2.6.1 Metodología para la construcción de ontologías: un enfoque basado en conceptos

Dada una perspectiva basada en conceptos respecto a la construcción de ontologías, Buiteelar *et al.* (2005) proponen una metodología compuesta por seis etapas (Ver figura 6).

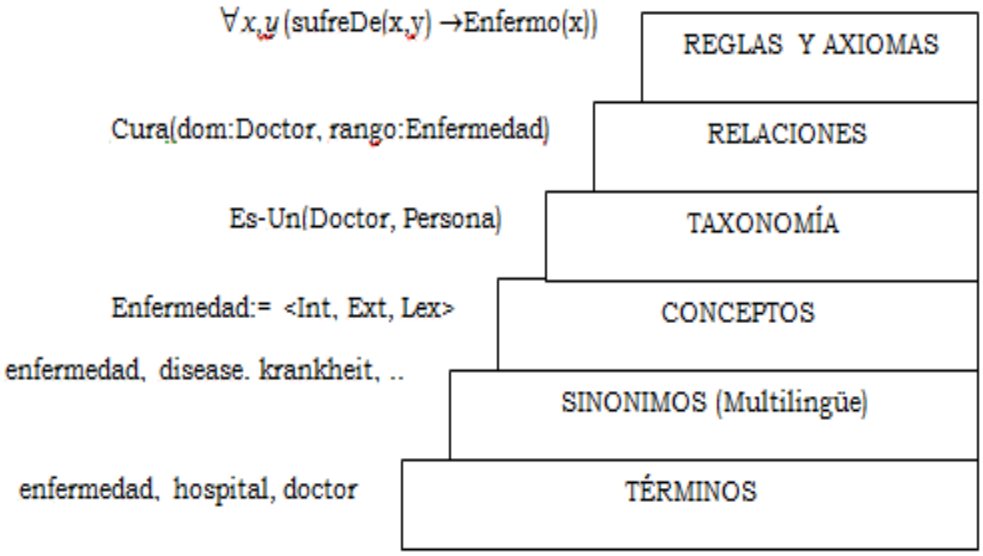

Figura 6. Metodología para el *aprendizaje* de ontologías, extraída de Buitelaar *et al.* (2005).

Esta metodología parte del análisis de textos, donde cada etapa, a grandes rasgos, consiste en lo siguiente:

Extracción terminológica

La primera fase consiste en la extracción de los términos o unidades léxicas usadas para denotar los conceptos relevantes del dominio. Un primer enfoque para hacer lo anterior es mediante el reconocimiento de patrones de

constitución de términos (enfoque lingüístico). Por ejemplo, los patrones más comunes de construcción de términos en español son:

- <NombreComún>: enfermedad, doctor, etc.
- <NombreComún><Adjetivo>: enfermedad degenerativa, infección urinaria, etc.
- <NombreComún><FrasePreposicional>: Síndrome de Down, Enfermedad de Crohn.

Algunos ejemplos de extractores terminológicos desarrollados bajo este enfoque son: LEXTER (Bourigault, 1994) y HEID (Heid, 1999). Por otro lado, el enfoque estadístico parte del supuesto de que si una frase nominal (FN) ocurre con frecuencia en un corpus, esto indica una probabilidad alta de que se trate de un término. Un ejemplo de extractor terminológico desarrollado bajo este enfoque es ANA (Enguehard, 1993). Finalmente, se han propuesto métodos híbridos que consideran una combinación de los enfoques lingüístico y estadístico. Ejemplos de este tipo de extractores son ACABIT (Daille, 2003), TermoStat (Drouin, 2003), TermExt (Barrón, 2007), y YATE (Vivaldi, 2001).

Extracción de sinónimos

La etapa de extracción de sinónimos plantea la identificación de términos que potencialmente denotan el mismo concepto en una o varias lenguas, siendo esto último de gran interés para la traducción automática de términos. Esta etapa involucra también la consideración de mecanismos de desambiguación del sentido de un término (WSD) en aras de determinar los sinónimos que serán extraídos.

De acuerdo con Buiteelar *et al.* (2005), en el contexto de la construcción de ontologías se ha explotado el hecho de que los términos ambiguos tienen significados muy específicos dentro de dominios restringidos, lo que permite un enfoque integrado para la desambigüación de sentido y la extracción de sinónimos. Entre los métodos o técnicas usados en esta fase están el *clustering* y la indexación de semántica latente, por mencionar algunos (Pinto *et al.* 2007; Guzmán *et al.,* 2009; Agirre y Soroa, 2007; Lesk, 1986; Purandare y Pedersen, 2004; Schütze, 1998; Sussna, 1993; Yarowsky, 1992).

Extracción conceptual

La etapa de extracción de conceptos es muy importante y por supuesto no libre de controversia. Como lo menciona Smith (2004), el problema es que en la mayoría de los trabajos sobre ontologías, bajo el enfoque basado en conceptos, no se ofrece una definición clara de qué es un concepto, por ello resulta difícil determinar y mucho más consensuar al respecto.

Buiteelar *et al.* (2005) proponen que la formación o inducción de un concepto debe proveer de lo siguiente:

- Una definición intensional, es decir, un conjunto de propiedades necesarias y suficientes para describir un concepto. Estas propiedades dan cuenta de las relaciones del concepto con otros. Por ejemplo, el concepto *perro*, en una definición analítica *perro es un animal*, se encuentra relacionado con *animal* en términos de una relación subordinado-superordinado (IS-A).
- Una extensión, es decir, un conjunto de instancias del concepto. Para el caso del concepto *Animal,* una definición por extensión podría estar representada por un conjunto como A= {gato, perro, vaca, león}, que son elementos que tienen algunas propiedades en común y que caen dentro de la categoría genérica *animal.*
- Un conjunto de etiquetas lingüísticas (términos) para denotar el concepto. Por ejemplo, para denotar el concepto *perro* tenemos los siguientes términos: can, chucho, tuso[14], etc.

La información conceptual, en los términos propuestos por los autores mencionados anteriormente, puede ser extraída automáticamente mediante un enfoque lingüístico (Rebeyrolle y Tanguy, 2000; Muresan y Klavans, 2002; Malaisé, 2005; Sánchez y Márquez, 2005; Storrer y Wellinghoff, 2006; Rodriguez, 2005; Saggion, 2004; Monachesi, 2007; Sierra *et al.*, 2008). Este enfoque consiste en identificar patrones que introducen un término y su definición correspondiente.

[14] Sinónimos extraídos de: http://www.wordreference.com/sinonimos/perro

Extracción de taxonomías

La cuarta etapa consiste en la extracción de taxonomías o jerarquías de conceptos. De acuerdo con Winston *et al.* (1987), las relaciones que estructuran el espacio conceptual de forma jerárquica son las relaciones IS-A y parte-todo. Sin embargo, en la literatura solo se ha utilizado para este fin la relación de hiponimia-hiperonimia.

Los enfoques utilizados para la extracción de taxonomías a partir de textos son básicamente los siguientes:

- Patrones léxico-sintácticos (Hearst, 1992; Charniak y Berland, 1999; Pantel y Pennacchiotti, 2006; Ortega *et al.*, 2007), para extraer relaciones de hiponimia y meronimia, respectivamente.
- Hipótesis distribucional de Harris. Este enfoque considera la distribución contextual para derivar jerarquías de términos mediante la aplicación principalmente de técnicas *clustering* (Pereyra, 1993; Faure *et al.*, 1998).
- Basado en la noción de subsunción de términos. En este enfoque se establece la generalidad relativa de dos términos a partir de considerar que un término t_1 es una subclase de t_2 si todos los contextos sintácticos en los que t_1 aparece son también compartidos por t_2 (Cimiano *et al.,* 2004).

Extracción de relaciones no taxonómicas

La quinta fase consiste en la extracción de relaciones no taxonómicas, es decir, relaciones donde no existen jerarquías entre los conceptos. La mayoría de los enfoques aplicados en esta tarea combinan análisis estadísticos y niveles algo complejos de análisis lingüístico, como por ejemplo, la explotación de la estructura sintáctica y de dependencias de una oración. En esta tarea, los elementos de una oración, que cubren una función sintáctica específica (sujeto, objeto), configuran roles temáticos de algún tipo (agente, paciente, tema) dependiendo del verbo y la construcción sintáctica del mismo. Por ejemplo, en la oración *el doctor cura enfermedades,* se tiene que el verbo (predicado) *cura* tiene dos argumentos: *doctor* y *enfermedad*, donde *doctor* cubre el rol de agente y *enfermedad* el de tema (Calvo, 2006; Segura *et al.*,

2006; Dorr *et al.*, 1994; Pugeault *et al.*, 1994).

Extracción de reglas de inferencia

Finalmente, la última etapa consiste en la extracción de reglas de inferencia. De acuerdo con el trabajo de Lin y Pantel (2001), es posible descubrir automáticamente un conjunto de reglas de inferencia a partir de textos. Un ejemplo de este tipo de reglas es: *Si X es autor de Y entonces X escribió Y*. En este trabajo se aplica un enfoque no supervisado que considera como base la hipótesis distribucional de Harris, aunque aplicada a rutas (paths*)* de árboles de dependencia derivados de un corpus con etiquetado sintáctico.

Los trabajos en esta última etapa son muy escasos, incluso para el inglés, por lo que esto deriva en un campo muy fértil para la investigación.

2.7 Resumen del capítulo

La adopción de la visión clásica como una meta-teoría para formular conceptos científicos ha evidenciado los problemas que se han señalado sobre este enfoque a lo largo de este capítulo: conceptos no consensuados y fracaso para especificar rasgos necesarios y suficientes (Smith y Medin, 1981). Por lo menos en Biología, existen una gran cantidad de conceptos biológicos que no han sido consensuados en términos de la clase a la que pertenecen. Por ejemplo, no existe acuerdo entre biólogos sobre si *Euglena* se debe clasificar como un *animal* o como una *planta*. Ante problemas como el anterior, ¿cómo determinar los rasgos de la *differentia* si la clase genérica no está bien definida? Esto sin duda deriva en cuestionamientos sobre la validez del enfoque clásico para efectos de clasificación biológica.

A pesar de lo anterior, sigue siendo una práctica común enunciar definiciones para dar cuenta del significado de conceptos. Existen otros esquemas de definición que prescinden de un *genus* y sólo enfatizan la función del concepto (definiciones funcionales), o bien las partes que constituyen una entidad (definición meronímica). Esta variedad de información respecto a un concepto léxico contribuye a configurar un significado más completo (Pustejovsky, 1991), por lo menos válido dentro de un dominio específico (Fillmore, 1975).

En disciplinas como la lexicografía computacional y la terminología, la

búsqueda de definiciones en corpus para elaborar diccionarios electrónicos o bancos de consulta terminológica, es una tarea que tiene ya una larga tradición y que ha rendido frutos importantes (Aguilar, 2009). La terminología o conjunto de unidades léxicas de un dominio tienen un significado generalmente único y posiblemente complejo dentro del dominio. En cierta forma, la terminología es la apariencia superficial presente en textos del conocimiento de un dominio. Debido a su baja ambigüedad y alta especificidad son también particularmente útiles para conceptualizarlo (Velardi *et al.*, 2001).

De acuerdo con la revisión de este capítulo, podemos esperar que en el análisis de enormes colecciones de textos de un dominio encontremos más de una definición del mismo concepto donde varíe la categoría asignada. En algunos casos, dichas categorías pueden considerarse como variantes léxicas, por ejemplo, en el caso de *inflamación* e *hinchazón* cuando se asignan a un concepto como *bursitis*. En otro escenario, las categorías pueden ser miembros de una jerarquía de categorías, como en el caso de *mamífero-animal* cuando se encuentran categorizando al concepto *ballena*. En estos casos, de acuerdo con los resultados de estudios de la psicología cognitiva, esperaríamos que, en dominios especializados, la categoría genérica fuese por lo menos el *genus* o una categoría subordinada. Por último, otro escenario, aún más complejo, y que se deriva de las fuertes críticas hechas a la visión clásica, es el caso de conceptos no consensuados, incluso en dominios específicos, como se mencionó al inicio de este apartado. Esta última situación derivaría en más de un hiperónimo asignado al concepto, con una o más *differentiae* acorde al *genus* asignado.

"(…) As for any other phenomenon in the world, the existence of paradigmatic semantic relations among words calls for some kind of explanation – or perhaps several kinds of explanation. Are these relations among words, or among the things the words represent? Are the relations arbitrary or rule based? Language specific or universal? A product of linguistic or general cognition?"

Semantic relations and the Lexicon. M. Lynne Murphy (2003)

Capítulo 3

Relaciones léxico-semánticas

En la actualidad, en la literatura sobre el tema, existe un uso indistinto de los términos relaciones léxicas y relaciones semánticas para denominar aquellas relaciones que se establecen entre las palabras o los conceptos que éstas representan. Cualquiera de estos dos términos hace alusión básicamente a las relaciones paradigmáticas: hiponimia-hiperonimia, antonimia, incompatibilidad, meronomia-holonimia y sinonimia. Una relación paradigmática es aquella que se da entre los miembros de un conjunto donde un elemento cualquiera del conjunto puede sustituir a otro en el contexto de una oración sin que esto afecte su gramaticalidad (Saussure, 1968).

En Croft y Cruse (2004) se presenta una breve discusión en el aspecto siguiente: la mayoría de los textos sobre semántica léxica señalan que las relaciones no se dan entre las palabras como tales, sino entre los conceptos que denotan las palabras, lo que resta conveniencia al término *léxico*. Por ejemplo, en algunos textos estas relaciones se denominan *relaciones de sentido* (Lyons, 1977) o *relaciones de significado* (Allan, 1986). Dado lo anterior, el planteamiento más consensuado parece ser que las relaciones son semánticas porque se dan entre los sentidos de las palabras. A pesar de que algunos están más a favor del término *semántico*, como es el caso de Lyons (1977), el mismo autor las ha considerado como propiedades estables de elementos léxicos. Por su parte, la comunidad de lingüística cognitiva no ha aportado mucho debate en torno a esta discusión.

Respecto al término *relación*, Murphy (2003) explica que el sentido de esta palabra, en el término *relación léxica o semántica,* es aquel denotado por las relaciones paradigmáticas. La palabra *relación* describe la membresía a un conjunto definible. De este modo, palabras como *azul* y *rojo* están relacionadas porque pertenecen al conjunto de palabras que refieren al color de un objeto, por lo que podríamos etiquetar esta relación como *color*.

Dada la falta de consenso entre utilizar léxico o semántico, se asumirá en este trabajo el término compuesto *relaciones léxico-semánticas* para denotar el tipo de relación en que se enfoca esta investigación.

3.1 Relaciones paradigmáticas y sintagmáticas

Respecto al significado de una palabra, Cruse (1986) señala que cada aspecto del significado se refleja en un patrón característico de normalidad (o anormalidad) en contextos gramaticalmente apropiados. Básicamente, las afinidades pueden ser de dos tipos, sintagmáticas o paradigmáticas. Las afinidades sintagmáticas se establecen vía la capacidad de asociación normal en un discurso y siempre presuponen una relación gramatical específica. Por ejemplo, en dichos términos, *el perro ladró* es una construcción normal. Una desafinidad se manifiesta por medio de una anormalidad que no viola restricciones gramaticales, por ejemplo, *los leones están gorjeando*. Por otro lado, una afinidad paradigmática entre dos palabras con la misma categoría gramatical es mayor mientras más congruente sean sus patrones de normalidad sintagmática. Así, por ejemplo, *perro* y *gato* comparten más contextos normales y anormales que, *perro* y un *poste de luz*:

- César alimentó el gato/perro/*poste de luz.
- El perro/gato/*poste de luz huyó.
- El *perro/*gato/poste de luz se dobló en el accidente.
- Pintamos el *perro/*gato/poste de luz.

3.2 Sentidos de palabras

Un elemento léxico puede tener más de un sentido o significado. En los diccionarios esta situación es muy evidente debido a que generalmente se ofrece más de un significado para una palabra. Por ejemplo, algunos de los significados para la palabra *perro* obtenidos del diccionario de la RAE, son los siguientes:

Perro[1], rra.

1. adj. coloq. Muy malo, indigno.

2. adj. *El Salv.* Dicho de una persona: Enojada, de mal genio.

Perro[2]

1. m. Mamífero doméstico de la familia de los Cánidos, de tamaño, forma y pelaje muy diversos, según las razas. Tiene olfato muy fino y es inteligente y muy leal al hombre.

2. m. U. por las gentes de ciertas religiones para referirse a las de otras por afrenta y desprecio.

3. m. Persona despreciable.

4. m. Mal o daño que se ocasiona a alguien al engañarle en un acuerdo o pacto.

5. m. desus. Hombre tenaz, firme y constante en alguna opinión o empresa. Era u. t. c. adj.

En el ejemplo anterior se muestra sólo una parte de los diferentes sentidos de la palabra *perro*, sin embargo, basta con este subconjunto para vislumbrar la amplia gama de significados atribuidos a la palabra. Dado lo anterior, si intentamos escudriñar en nuestro lexicón mental y dar cuenta de cómo representamos nuestro vocabulario, podríamos llegar a la conclusión de que la representación es similar a la de un diccionario. A este respecto, Murphy (2003) señala que esta visión no es la adecuada debido a que ambos ofrecen información diferente. Por ejemplo, el lexicón mental debe registrar que *contento* no se usa en una posición prenominal porque incurre en una agramaticalidad; esta información no la registra generalmente un diccionario. Aunado a lo anterior, el énfasis de los diccionarios es listar todos los significados de las palabras, sin embargo, esto no es posible dado el uso semántico ilimitado que podemos hacer de ellas. Por otro lado, a diferencia de los diccionarios, el lexicón mental no puede hacer una separación arbitraria del significado definicional y del enciclopédico, ni son sus divisiones de sentidos semejantes a aquellas de los diccionarios.

Para Fellbaum (1998) es frecuente que las personas hagan una distinción entre conocimiento de una palabra (léxico) y conocimiento del mundo (o enciclopédico). Los diccionarios y las enciclopedias reflejan precisamente esta distinción. Los diccionarios ofrecen el conocimiento de las palabras y las enciclopedias el conocimiento del mundo; sin embargo, los límites de los dos son difusos. Por ejemplo, la mayoría de las personas estaría de acuerdo en que *golpear a alguien* es un acto hostil y que este *saber* es parte de nuestro conocimiento del mundo. Por otro lado, saber que el verbo *golpear* es un verbo fuerte, que tiene una relación de sinonimia con el verbo *atacar*, y

que toma un argumento directo constituye conocimiento de la palabra. Así, dado lo anterior, la conclusión que parece más razonable es que la comprensión del significado y el uso de una palabra no se limitan a un solo tipo de conocimiento.

Con un enfoque más práctico, Hirst (2004), al hablar de las diferencias entre ontologías y lexicones, señala que los sentidos de las palabras tienden a ser objetos difusos, en ocasiones también con límites difusos. Por ejemplo, la palabra en inglés *open* tiene muchos sentidos que se traslapan: *unfolding*, *expanding*, *revealing*, *moving to an open position*, *marking openings in*, etc., por tanto, la separación en sentidos discretos es prácticamente imposible. Dado lo anterior, se plantea que los sentidos de las palabras se derivan, crean o modulan en cada contexto de uso (Croft y Cruse, 2004). Hirst (2004) menciona que a pesar de los problemas que implican la multiplicidad de sentidos, para aplicaciones prácticas, puede resultar conveniente asumir que el sentido de una palabra es una categoría. Por ejemplo, en el contexto de una ontología específica, cada sentido de palabra se representa simplemente como un apuntador a un concepto o categoría dentro de la ontología. En algunos dominios técnicos esto podría ser muy conveniente, sin embargo, en algunas ocasiones podría no ser la mejor decisión debido a que todos los problemas derivados de la categorización seguirían vigentes.

3.3 La organización de un espacio conceptual

Uno de los supuestos que ha guiado la investigación en semántica es que las palabras denotan conceptos, o unidades de significado (Croft y Cruse, 2004). Si partimos de este supuesto, entonces, al intentar proveer de estructura a un conjunto de unidades léxicas dentro de un dominio, estaremos también imprimiendo estructura en un espacio conceptual y con ello derivando conocimiento. En este sentido, la semántica léxica, con una visión estructuralista, considera varios tipos de relaciones: sinonimia, antonimia, hiponimia-hiperonimia, meronimia-holonimia. Estas relaciones serán descritas con más detalle en la sección 3.4.

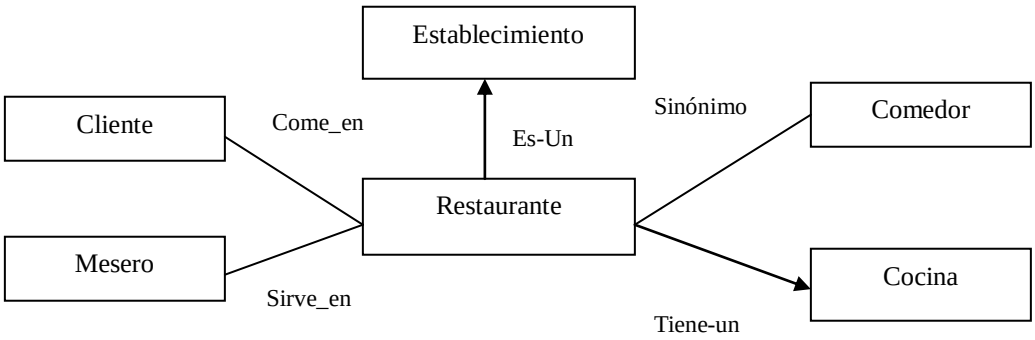

Figura 7. Concepto restaurante y sus conceptos relacionados.

Las relaciones paradigmáticas sin duda contribuyen a organizar conceptos de acuerdo con las relaciones que establecen con otros conceptos, sin embargo no abarcan la amplia gama de asociaciones que pueden existir. En la sección 3.1, Cruse señala que cada aspecto del significado de una palabra se refleja en patrones de normalidad (anormalidad) sintagmáticas y paradigmáticas, lo que evidencia que lo paradigmático no es suficiente. Con esto en mente, Fillmore propuso su *semántica de marcos*. Este enfoque se inspiró en la noción de marcos (*frames*) de Minsky y se enfocó a dar cuenta de que existen otras relaciones que se dan entre los conceptos que no son precisamente paradigmáticas y que se explican simplemente porque están ligados por la experiencia. Un ejemplo clásico es el concepto *restaurante.* Visto desde el plano de las relaciones paradigmáticas, un restaurante es un establecimiento de servicio (hiponimia-hiperonimia), restaurante-comedor (sinonimia), cocina-restaurante (meronimia-holonimia). Sin embargo, existen otros conceptos tales como *cliente*, *mesero*, *orden*, *factura*, *menú*, que no se vinculan por ninguna de las relaciones anteriores, y se encuentran estrechamente ligados con el concepto *restaurante* (Véase la figura 7).

3.4 Tipos de relaciones léxico-semánticas paradigmáticas

Las relaciones léxico-semánticas más exploradas en la literatura, aunque con niveles diferentes de exploración, son la hiponimia-hiperonimia, antonimia, incompatibilidad, sinonimia y meronimia-holonimia. La siguiente descripción se hace principalmente con base en los trabajos de Cruse (1986), Krifka (1998), Winston *et al.* (1987) y Murphy (2003). Se considera pertinente mostrar las

aplicaciones de este tipo de relaciones para tener una idea clara de su importancia. Para lograr este fin, se describe la organización de cada una de las relaciones en un proyecto de base de datos léxica denominado WordNet[15], que fue fundado en 1985, en el Cognitive Science Laboratory de la Universidad de Princeton.

Algunas cuestiones generales sobre WordNet

WordNet es una base de datos léxica para el inglés con nombres, verbos, adjetivos y adverbios que se agrupan en conjuntos de sinónimos cognitivos (synsets), donde cada uno expresa un concepto distinto (Fellbaum, 2005). WordNet ha sido utilizado en varias tareas de procesamiento de lenguaje natural (PLN): desambiguación de sentidos, traducción máquina, etc. Actualmente, el modelo WordNet se sigue replicando para otros lenguajes, como es el caso de EuroWordNet. El éxito de WordNet se debe, en parte, al intento por construir un lexicón computacional de tamaño humano, mientras que la mayoría de los proyectos de PLN se construyen y prueban con lexicones limitados. El grupo que desarrolla WordNet plantea que manipular lexicones grandes revela problemas y patrones específicos que no se observarían en lexicones pequeños.

Para efectos de la manipulación del tamaño, WordNet se divide en tres lexicones que organizan nombres, verbos y modificadores, respectivamente. Cada lexicón opera con sus propios principios de organización. Debido a que las palabras se organizan dentro de su propia categoría gramatical, las relaciones que se representan en WordNet, son principalmente paradigmáticas. Incluso para los verbos se plantea una organización paradigmática, lo que va más acorde con la conexión psicolingüística de la que se deriva este modelo.

3.4.1 Sinonimia

La sinonimia implica términos refiriendo a un mismo concepto. Dos expresiones a y b de un lenguaje L se consideran sinónimos si y solo sí tienen el mismo significado. La sustitución entre sinónimos no cambia las condiciones de verdad de las construcciones más grandes (oraciones,

[15] Para consultas al lexicón WordNet: http://wordnet.princeton.edu/

proposiciones, etc.) en que son usados.

La sinonimia tiene la propiedad de ser simétrica y por tanto no jerárquica. Sea R la relación binaria denotada por el símbolo "=" sobre un lenguaje L:

$$\text{Si } a, b \in L \text{ son sinónimos, entonces } aRb \rightarrow bRa$$

Aplicación de la sinonimia

En WordNet, las palabras se agrupan en conjuntos de sinónimos cognitivos (synsets), donde cada conjunto expresa un sentido diferente, es decir, si una palabra tiene más de un sentido, se representa en más de un synset. Los synsets se relacionan con otros por medio de otras relaciones paradigmáticas. Las relaciones que se incluyen dependen de la categoría gramatical de la palabra. Por ejemplo, la figura 8 muestra el resultado de la búsqueda de la palabra *open* con la herramienta WordNet. La palabra *open*, como nombre, tiene cuatro sentidos diferentes y las relaciones que se incluyen en esa categoría son principalmente de hiperonimia. Por otro lado, para el caso de la misma palabra con categoría de verbo, existen 11 sentidos (ver figura 9). Para el caso de verbos, las relaciones que se tienen son de troponimia, causalidad, antonimia y marco oracional.

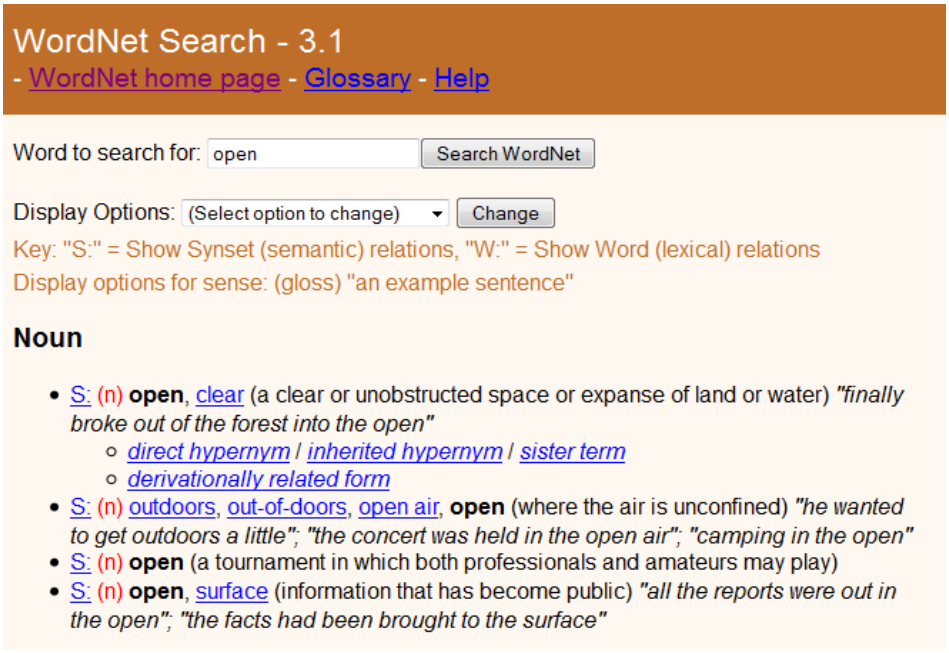

Figura 8. Sentidos de la palabra *open* (nombres).

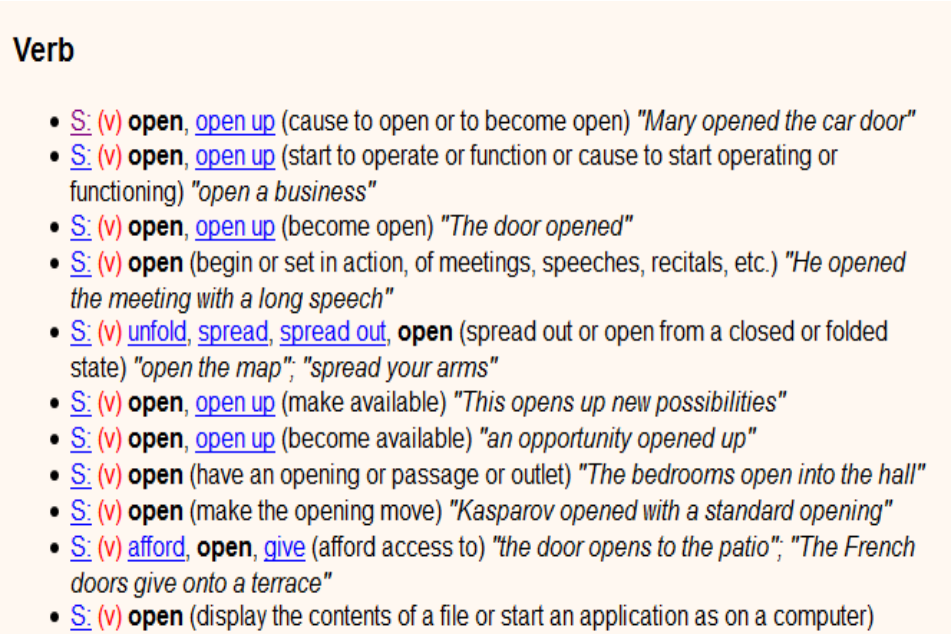

Figura 9. Sentidos de la palabra open (verbos).

A manera de comentario sobre la discusión del término apropiado para denominar estas relaciones, los creadores de WordNet plantean una distinción entre relaciones léxicas, que incluyen formas de palabra y significados, así como relaciones conceptuales, que incluyen sólo significados (Miller *et al.*, 1990; Fellbaum, 1998). La diferenciación anterior se hace considerando que entre *synsets* se dan las relaciones conceptuales, y las relaciones léxicas se dan entre palabras específicas. La sinonimia se trata siempre como una relación léxica debido a que generalmente las palabras en un *synset* se corresponden sólo con un concepto. A pesar de lo anterior, no siempre es claro cuando se hace esta distinción. Esta situación reafirma también la ambigüedad y falta de consenso sobre cómo etiquetar este tipo de relaciones en la práctica.

3.4.2 Antonimia

Si dos expresiones a y b no pueden aplicar a la misma entidad, se pueden considerar incompatibles. Un caso de incompatibilidad son los términos complementarios, ejemplos de ello son: abierto/cerrado, vivo/muerto, falso/verdadero. Si un caso pertenece a alguno de ellos, no puede pertenecer al otro. Otro tipo de incompatibilidad es la antonimia, donde la mayoría de los

ejemplos cae en la categoría gramatical de los adjetivos: barato/caro, bajo/alto, rápido/lento. Del mismo modo, también existen ejemplos de pares de verbos: levantar/caer (Krifka, 1998).

El contraste entre términos complementarios y la antonimia es que en el primero un caso particular pertenece a cualquiera de dos posibles términos complementarios. En el segundo, los pares de adjetivos pueden adquirir formas graduables debido a que tienen formas comparativas (más barato que) y superlativas (el más barato).

Aplicación de la Antonimia

WordNet distingue antonimia directa (léxica) e indirecta (conceptual) y organiza el lexicón de adjetivos en synsets alrededor de adjetivos focales. Por ejemplo, *wet* y *dry* son adjetivos focales, relacionados por antonimia directa. *Damp, moist*, y *humid* se relacionan con *wet* mediante ligas sinónimas. Del mismo modo, *arid, sere,* y *parched* se ligan a *dry* como sinónimos. *Arid/humid* o *moist/parched* son antónimos indirectos porque se oponen cuando se considera la antonimia directa entre *wet* y *dry* (Véase figura 10).

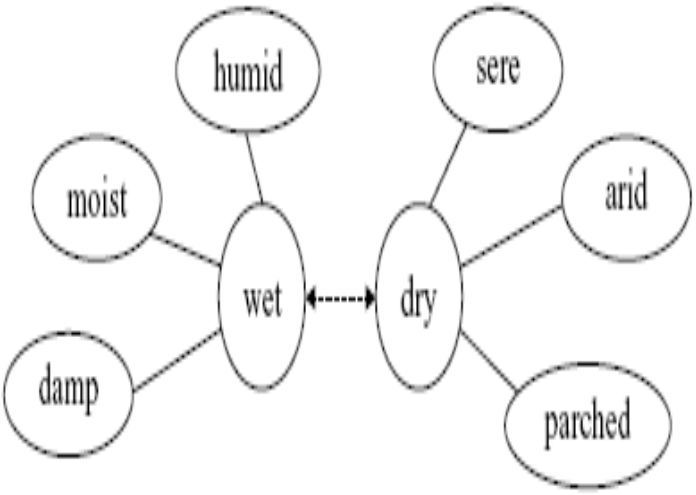

Figura 10. Adjetivos descriptivos en WordNet, extraído de Murphy (2003).

Algunas consideraciones importantes sobre adjetivos relacionales

En WordNet, los *synsets* de adjetivos se organizan en una red paradigmática de antónimos. Sin embargo, existen muchos adjetivos relacionales que no

tienen antónimos obvios. Así, este último tipo de adjetivos se representa con apuntadores a nombres con los que comparte contenido semántico (Murphy, 2003).

3.4.3 Hiponimia-hiperonimia

La relación de hiponimia/hiperonimia se considera como una de las relaciones estructurales más importantes en el lexicón (Lyons, 1968; Cruse, 2002), y representa la relación léxico-semántica que se ha estudiado más en la comunidad computacional.

Dado lo mencionado en el párrafo anterior, la pregunta es: ¿por qué la relación de hiponimia es tan importante en muchos modelos del lexicón y en inteligencia artificial?, esta es una pregunta que se responde casi por sí misma: por su naturaleza inferencial, particularmente por sus propiedades de implicación; su importancia en la definición, y su relevancia para restricciones selectivas en gramática.

En el aspecto inferencial, si una expresión a es un hipónimo de una expresión b, todo lo que cae bajo b también cae bajo a, en este caso b es denominado hiperónimo. Por ejemplo, *perro* es un hipónimo de *animal* y, en la relación inversa, *animal* es un hiperónimo de *perro*. Por tanto, es posible argumentar lo siguiente:

> Una expresión a es un hipónimo de una expresión b si cada oración declarativa donde se encuentra a [...a...], por ejemplo, en la oración: *este perro es muy tranquilo*, implica la oración donde a se remplaza por b [...b...], pero donde b no es implicado por a: *Este animal es muy tranquilo*.

La condición entre paréntesis proporciona una noción estricta de hiponimia, es decir, aquella que excluye la sinonimia. Se habla por tanto de relaciones de implicación unilaterales (símbolo para denotar relaciones de implicación unilaterales |→):

Pedro se comió una manzana. |→Pedro se comió una fruta.

Dado lo anterior, podemos deducir que la relación de hiponimia-hiperonimia es antisimétrica y por tanto jerárquica, es decir, impone un orden en la estructura conceptual. Sean a y b dos expresiones del lenguaje L y R la relación "<" (menor que), si a y b mantienen una relación de hiponimia-

hiperonimia, entonces se cumple lo siguiente:

$$\forall a, b \in L, aRb \rightarrow b\neg Ra$$

Por otra parte, las definiciones clásicas (Aristotélicas) se basan también en la relación de hiponimia-hiperonimia. Estas definiciones, que son comunes en diccionarios estándares y en textos de dominios especializados, consisten de un *genus* (hiperónimo) y una o más *differentiae* que distingue el concepto definido (hipónimo) de otros miembros del hiperónimo, como se mencionó en el capítulo 2.

Finalmente, en términos gramaticales, las restricciones selectivas en el objeto de un verbo pueden parafrasearse en términos de un hiperónimo y sus hipónimos. Así, todos los hipónimos pueden seleccionarse también como potenciales objetos. Por ejemplo, el verbo *beber* selecciona un hiperónimo como *bebidas* y, a partir de esto, todos sus hipónimos (agua, cerveza, jugo, etc.) pueden ser objetos del mismo verbo.

Tipos de hiponimia

Para Croft y Cruse (2004) las relaciones de hiponimia-hiperonimia son de inclusión, particularmente de dos tipos: *hiponimia simple* y *taxonimia*. La hiponimia simple podemos representarla lingüísticamente como *X es un Y*. Por su parte, la taxonimia puede ejemplificarse vía la construcción lingüística *X es un tipo/clase de Y*. Esta última relación discrimina más que la hiponimia simple, y deriva generalmente en una relación taxonómica. Aunado a lo anterior, Croft y Cruse señalan que en muchos casos donde un buen hipónimo no es un buen taxónimo de un hiperónimo, existe una definición directa del hipónimo en términos del hiperónimo más un rasgo semántico simple:

Semental = Caballo macho

Otro caso como el anterior donde cada uno de los adjetivos es un modificador del nombre *cuchara* con un rasgo simple y que muestran poca o nula utilidad para elaborar una clasificación del hiperónimo es el siguiente:

C (cuchara, w) = (redonda, grande, profunda…)

Por otro lado, podemos encontrar también modificadores adjetivos relacionales que proporcionen perspectivas de división útiles debido a que

enfatizan un aspecto importante del concepto, en el siguiente caso, el aspecto funcional:

C (cuchara, w) = (sopera, cafetera, pastelera,…)

Finalmente, es posible también encontrar modificadores de frase preposicional que enfaticen también el aspecto funcional y que sean producto de una variación de permutación (Vivaldi, 2001):

C (cuchara, w) = (de sopa, de café, de pastel…)

La pregunta que surge aquí es: ¿los rasgos conceptualmente simples podrían ser indicativos de hipónimos no relevantes? Si la respuesta es afirmativa, entonces es necesario discernir si tal relación muestra rasgos conceptualmente simples o complejos, de modo que esto ayude a ubicar hipónimos que expresen valoraciones generales, *versus* aquellos que configuren una red conceptual jerarquizada subyacente en un dominio especializado.

Aplicación de la hiponimia-hiperonimia

En WordNet, tanto los nombres como los verbos se representan en estructuras jerárquicas, mientras que los adjetivos tienen una representación no jerárquica. Este planteamiento de estructura variada se debe a las diferentes relaciones presentes en cada una de las tres categorías sintácticas y a las prioridades que se les dan en cada lexicón. En lo que respecta a los nombres, en WordNet se organizan mediante hiponimia-hiperonimia, antonimia, y meronimia-holonimia. La hiponimia-hiperonimia sirve como principio de organización, ya que todos los nombres participan en alguna relación de inclusión. Con esta meta de estructuración, se proponen 25 clases en el nivel más alto de la jerarquía (ver figura 11), entre las que se encuentran: *animal*, *comunicación*, *ubicación*, *relación*, y *sustancia* (Miller, 1998). Relaciones como la hiponimia y la hiperonimia se consideran conceptuales porque relacionan *synsets* en lugar de palabras.

{act, activity}	{food}	{possession}
{animal, fauna}	{group, grouping}	{process}
{artifact}	{location}	{quantity, amount}
{attribute}	{motivation, motive}	{relation}
{body}	{natural object}	{shape}
{cognition, knowledge}	{natural phenomenon}	{state}
{communication}	{person, human being}	{substance}
{event, happening}	{plant, flora}	{time}
{feeling, emotion}		

Figura 11. Categorías en el nivel más alto de la jerarquía de Nombres en WordNet. Extraído de Fellbaum (1998).

Por otro lado, para el caso de los verbos, existe también una estructuración jerárquica, donde se consideran varios tipos de implicación (ver Fellbaum, 2005, para más detalle). En este punto, surge la duda respecto a la organización paradigmática de los verbos. Si se supone que los verbos están más asociados con contextos sintagmáticos, ¿cómo es posible su organización paradigmática? En este sentido, Fellbaum (1998) menciona que el soporte básico para esta organización son los errores de sustitución. *Grosso modo*, esto consiste en analizar los verbos que los hablantes sustituyen erróneamente por otros. Por ejemplo, los dos pares de verbos *ask-tell* y *go-come*. Estos pares de verbos son del mismo dominio semántico y tienden a seleccionar sujetos relacionados semánticamente.

Finalmente, un punto importante que resalta Miller (1998), y que se relaciona con lo mencionado en el capítulo 2, es que la representación jerárquica se basa en el modelo clásico de categorización. Como se explicó anteriormente, la visión clásica de categorización plantea problemas debido a lo difuso que pueden resultar los límites de las clases donde cada una de las entidades del mundo puede ubicarse, además del efecto de *tipicalidad* que hace a algunos miembros de la clase más prototípicos o representativos que otros. En el caso de WordNet, Miller señala que la estructura jerárquica del lexicón de nombres parece ajustarse a hechos lingüísticos y ello le da pie a sugerir, a falta de una mejor explicación, que las representaciones prototípicas de categorías y la representación jerárquica del lexicón de nombres coexisten.

3.4.4 Meronimia-holonimia

La relación léxico-semántica de meronimia-holonimia, o más coloquialmente conocida como parte-todo, es una relación entre partes de cosas u objetos y los todos que las contienen (Winston *et al.*, 1987). Los ejemplos más intuitivos de esta relación son las partes del cuerpo: brazo, pierna, tronco y cabeza, por mencionar algunas (véase figura 12).

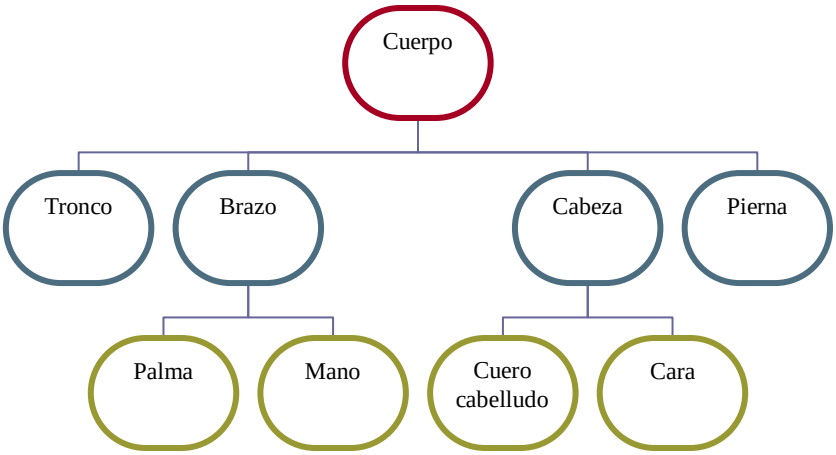

Figura 12**.** Partes del cuerpo humano.

De acuerdo con Winston *et al.* (1987), las relaciones meronímicas estructuran el espacio conceptual de forma jerárquica. Este tipo de relación tiene las propiedades de ser transitiva, antisimétrica y antireflexiva. Sea R la relación caracterizada por la expresión lingüística "parte de" y a, b, c ∈ D términos/conceptos de un dominio específico, entonces se tiene que:

- Transitiva: si aRb y bRc, entonces aRc
- Antirreflexiva: a¬Ra
- Antisimétrica: si aRb, entonces b¬Ra

Taxonomía de relaciones de meronimia

Uno de los trabajos más citados en la literatura sobre el estudio de la relación léxico-semántica meronimia-holonimia es el desarrollado por Winston *et al.* (1987). El resultado de la investigación de estos autores consiste en una taxonomía sobre diferentes tipos de meronimia. La taxonomía derivada de este trabajo se utilizó para explicar casos de intransitividad en silogismos merológicos y silogismos cuyas premisas expresan diferentes relaciones de

inclusión.

Aunque existen varias formas de expresar relaciones parte-todo, el trabajo tomado como base en esta sección se enfoca principalmente en la expresión "parte de" y sus variantes: *X es parte del Y*, *X es en parte de Y*, *Xs son parte de las Ys*, *X es una parte de Y*, *las partes de un(a) Y incluyen las Xs, Zs…* y son justamente estas expresiones características de la relación de meronimia las que dan lugar a una mejor comprensión de la taxonomía propuesta.

Una de las razones principales para pensar que existen diferentes tipos de relaciones meronímicas se deriva del denominado *criterio del argumento común.* Una forma de determinar que dos relaciones léxico-semánticas son diferentes es encontrar un caso en el que ambas apliquen al mismo concepto, pero contesten preguntas diferentes con respecto a él. Por ejemplo, un *canario* es un pájaro (hiponimia-hiperonimia), tiene alas (meronimia) y es de color amarillo (atribución). Si diferentes tipos de predicados pueden aplicar a un mismo concepto, decimos que existe un *argumento común.* Para el caso de la expresión evaluada: *parte de* y el argumento común *bicicleta*, tenemos las siguientes relaciones expresadas también por el término *parte*: *Las ruedas son parte de las bicicletas*, *las bicicletas son en parte de aluminio*, etc.

El *criterio del argumento común* soporta la distinción entre dos tipos de relaciones meronímicas: componente-objeto integral (pedal-bicicleta) y materia-objeto (aluminio-bicicleta). Sin embargo, la división es adecuada sólo si consideramos objetos físicos o sólidos. Por tanto, la clasificación debe tomar en cuenta los usos de partes respecto a colecciones, masas, actividades y áreas. En consideración con lo anterior, Winston *et al.* (1987) proponen una taxonomía de relaciones meronímicas que considera para diferenciarlas tres criterios importantes: si la parte tiene una función con respecto a su todo, si sus partes son idénticas a otras partes del mismo todo y si son separables o no del todo.

Un ejemplo de aplicación de los tres criterios es el siguiente: en la relación componente-objeto integral *llanta-carro,* la relación es funcional, una *llanta* no es idéntica a otras partes del mismo todo y es separable del todo. La tabla 7 muestra los seis tipos de relaciones, ejemplos y los criterios

mencionados (extraída y traducida de Winston *et al.,* 1987).

Tabla 7. Taxonomía de relaciones de meronimia-holonimia con criterios para discernir entre el tipo de relación.

Relación	Ejemplos	Criterios de relación		
		Funcional	**Homeomería**	**Separable**
Componente/Objeto integral	Pedal-Bicicleta	+	-	+
Miembro/Colección	Árbol-Bosque	-	-	+
Materia/Objeto	Aluminio-bicicleta	-	-	-
Porción-Masa	Rebana-Pastel	-	+	+
Rasgo-Actividad	Pagar-Comprar	+	-	-
Lugar-Área	Oasis-Desierto	-	+	-

Aplicación de la Meronimia-holonimia

WordNet contempla los siguientes tres tipos de meronimia que se incluyen en la taxonomía de Winston *et al.* (1987): componente-objeto integral (pierna-cuerpo), Miembro-colección (pariente-familia), y materia-objeto (aluminio-bicicleta). Debido a que no todos los nombres entran en una relación de meronimia, ésta se incluye sólo si aplica en un caso particular.

3.5 Las relaciones jerárquicas y los formalismos de representación

Antes de dar un panorama general sobre los formalismos de representación de relaciones jerárquicas, consideramos importante señalar que, en la literatura sobre extracción de relaciones léxico-semánticas, el único tipo de relación que ha sido considerada como jerárquica es la relación de hiponimia-hiperonimia. Sin embargo, de acuerdo con Winston *et al.* (1987), dadas las propiedades de ser transitiva, antisimétrica y antireflexiva, la relación parte-todo también estructura el espacio conceptual de forma jerárquica. De acuerdo con el modelo de WordNet, todos los nombres entran en alguna relación de inclusión de clase, lo que no sucede con las relaciones meronímicas.

El origen del interés por las taxonomías de conceptos

Poesio (2005) menciona que el interés de la comunidad de Inteligencia Artificial en taxonomías o jerarquías de conceptos inició con el trabajo sobre

redes semánticas de Quillian (1968), y la propuesta posterior de Minsky sobre representaciones basadas en marcos (Minsky, 1975). Las propuestas de estos dos investigadores introdujeron representaciones que no estaban basadas en lógica y que eran más intuitivas porque consistían en enfoques más cognitivos, por ejemplo, las estructuras de redes se derivaron de experimentos sobre recuperación de memoria humana. Estos enfoques de representación permitieron implementaciones de inferencias más fáciles y eficientes, tales como herencia y similitud, por medio de algoritmos de recorrido en árboles como *propagación de la activación*. La representación vía estos formalismos estaba cimentada en el uso de interfaces gráficas para modelar el conocimiento por medio de estructuras de datos *ad hoc*, donde el razonamiento se realizaba también mediante procedimientos *ad hoc* que manipulaban estas estructuras.

Modelos de red semántica

En 1968, Quillian propuso una técnica pionera denominada redes semánticas para el modelado de objetos. La meta de este modelo era caracterizar el conocimiento y el razonamiento de un sistema mediante estructuras cognitivas en forma de red. Este tipo de representaciones se llevaban a cabo generalmente mediante una estructura de grafos de datos donde los conceptos son los nodos y los arcos o enlaces representan las relaciones entre ellos (ver figura 13). En un modelo de red semántica no sólo se representan relaciones de inclusión de clase (IS-A), sino también relaciones de atribución, posesión, meronímicas y sintagmáticas, como puede observarse en la figura 13.

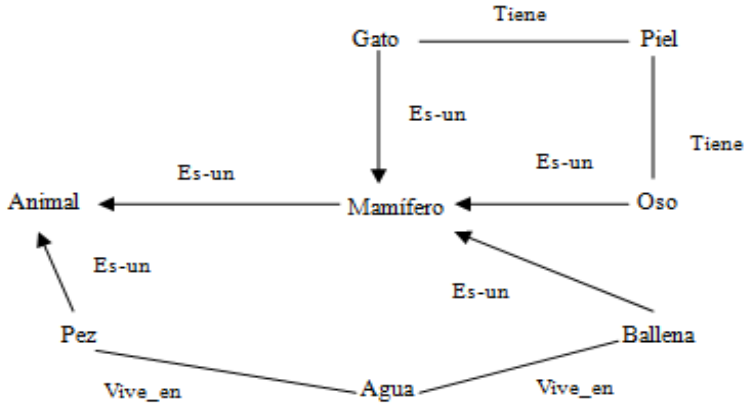

Figura 13. Red semántica.

Sistemas de marcos

Minsky (1975) fue el primero en proponer el enfoque de *marcos* para la representación del conocimiento. Este enfoque tenía objetivos similares a aquellos de las redes semánticas, sólo que se basaron en la noción de *marco* como un prototipo y en la capacidad de expresar relaciones entre *marcos*. Un marco es una colección de atributos que define el estado de un objeto y su relación con otros marcos u objetos (ver Figura 14). En Inteligencia Artificial, los marcos se denominan representaciones *slot-and-filler*. En primera instancia, los *slots* hacen alusión a los valores de los datos, y los *fillers* a los procedimientos que cambian los valores de estos datos u operan directamente con ellos para lograr un fin específico. La construcción de sistemas con marcos conectados entre sí es la forma más común de representación, lo que permite establecer un mecanismo de herencia para activar y actualizar los datos de los campos.

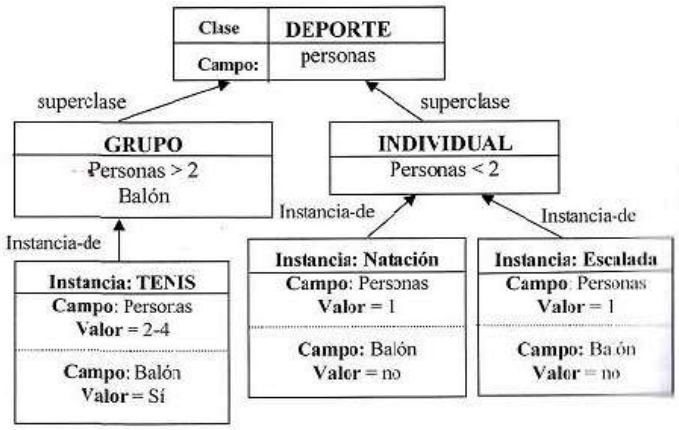

Figura 14**.** Representación del concepto *deporte* mediante marcos.
Extraída de Pajares y Santos (2006)

Una visión conjunta de los enfoques no basados en lógica

Aunque existen diferencias entre el modelo de red semántica y los sistemas de marcos, ambos tienen en común su origen cognitivo. Estas dos representaciones pueden considerarse como estructuras de red enfocadas a la representación de conjuntos de individuales y sus relaciones (Poesio, 2005).
Tomando como base sus orígenes cognitivos, los sistemas basados en red se consideraron más atractivos y más eficientes que los sistemas basados en

lógica. Desafortunadamente no fueron del todo satisfactorios debido a la falta de precisión en su caracterización semántica. Como consecuencia de lo anterior, el comportamiento de los sistemas que se implementaron con estos formalismos no era estable.

A raíz de la falta de caracterización semántica de los enfoques basados en red y de la importancia de explotar la noción de estructura jerárquica por los beneficios obtenidos en términos de la facilidad de representación y en eficiencia de razonamiento, se iniciaron trabajos para dotar a estos formalismos de una semántica más formal. Un primer logro en torno a este objetivo fue el descubrimiento de que se podía proveer de semántica a los *marcos* tomando como base la lógica de primer orden. En términos concretos, hacer lo anterior implicaba caracterizar los elementos básicos de la representación como predicados unarios y binarios. Por un lado, los predicados unarios denotaban conjuntos de individuales, y por el otro, los predicados binarios representaban las relaciones entre individuales. A pesar de este avance en la concepción del problema y posible solución, dicha caracterización no capturó las restricciones de las redes semánticas y la de los *marcos* con respecto a la lógica. De acuerdo con Brachman y Levesque (1985), aunque la lógica es la base natural para especificar significado para estas estructuras, resultó que las redes semánticas y los *marcos* no requerían de todo el aparato de lógica de primer orden, sino sólo una parte de él. La parte más interesante de este descubrimiento fue que las formas más comunes de razonamiento usadas en enfoques basados en estructuras podían lograrse por medio de técnicas de razonamiento especializadas, sin requerir de probadores de teoremas de lógica de primer orden.

Lógicas de descripción

Las lógicas de descripción es el nombre más reciente de un conjunto de formalismos de representación para modelar el conocimiento de un dominio (Jurafsky y Martin, 2009). A grandes rasgos, el procedimiento para modelar un dominio vía este formalismo consiste en definir los conceptos relevantes del dominio (su terminología), y después usar estos conceptos para especificar propiedades de objetos e individuales que ocurren en el dominio.

De acuerdo con Baader *et al.* (2003), la lógica de descripción se deriva de los enfoques de redes de herencia estructurada que se introdujeron para superar las ambigüedades de los primeros modelos de red semántica y marcos. Uno de los rasgos que distingue este formalismo de sus antecesores es precisamente que está equipado con una semántica formal y basada en lógica. Otro de los rasgos característicos es el énfasis en el razonamiento como un servicio nuclear: el razonamiento permite inferir conocimiento implícito del proporcionado explícitamente en la base de conocimientos. Las lógicas de descripción soportan patrones de inferencia que ocurren en muchas aplicaciones de sistemas de procesamiento de información inteligente, y que los humanos usan para estructurar y entender el mundo: la clasificación de conceptos e individuales.

La clasificación de conceptos determina las relaciones de subconcepto/superconcepto de una terminología y permite además estructurar la terminología como una jerarquía de conceptos. Esta jerarquía proporciona información útil sobre la relación entre conceptos y puede utilizarse para aumentar la velocidad de otros servicios de inferencia. Por su parte, la clasificación de individuales determina si un individuo particular es siempre una instancia de un concepto. Aunado a lo anterior, proporciona información útil sobre las propiedades de un individual.

Los problemas de decidibilidad y complejidad de la inferencia dependen del poder expresivo del lenguaje de lógica de descripción implementado. Es decir, si el lenguaje es muy expresivo es probable que se presenten problemas de inferencia de complejidad alta, o incluso llegar al punto de ser indecidibles. Por otro lado, lenguajes poco expresivos, asumiendo que tienen procedimientos de razonamiento eficientes, podrían no ser suficientemente expresivos para representar los conceptos importantes de una aplicación específica. Actualmente las investigaciones sobre este tipo de formalismos van encaminadas a lograr un equilibrio entre expresividad y complejidad de razonamiento.

3.6 Resumen del capítulo

Las relaciones léxico-semánticas representan un enfoque de la semántica estructural para organizar conceptos. Dicha organización contribuye a

configurar el significado de un concepto léxico (Cruse, 1986; Pustejovsky, 1991; Buitelaar *et al.*, 2005). Por ejemplo, la estructura Qualia propuesta por Pustejovsky para describir el significado de un elemento léxico consta de 4 roles básicos, dos de los cuales se corresponden con relaciones de hiponimia-hiperonimia (formal) y parte-todo (constitutivo). Aunado a lo anterior, se proponen dos roles más: uno agentivo que da cuenta del origen, y otro télico, que hace énfasis en la función del objeto. Por ejemplo, una descripción semántica mínima para el nombre *novela* podría incluir los siguientes valores para cada uno de los roles:

Novela (x):

Formal: libro(x), disco(x)

Constitutivo: Narrativa(x)

Télico: Leer(T, y, x)

Agentivo: artefacto(x), escribir(T,*z*,x)

De acuerdo con Pustejovsky (1991) la información anterior estructura nuestro conocimiento básico respecto al objeto *novela*: es una narrativa; su forma más común es la de un libro; para el propósito de lectura (con un tipo de evento de transición), y es un artefacto creado mediante un evento de transición de escritura.

Respecto al rol télico, consideramos importante mencionar aquí que Smith y Medin (1981) señalan que los experimentos en psicología sesgaron la consideración hacia rasgos estructurales (forma, tamaño, color, etc.) para definir propiedades necesarias y suficientes de un objeto o entidad. Sin embargo, esto no excluye que otro tipo de rasgos más abstractos, por ejemplo, funcionales, se puedan considerar para conformar la descripción unitaria de un concepto desde la perspectiva clásica. Los mismos autores mencionan que tal vez uno de los fracasos de esta incapacidad para enunciar esta descripción unitaria es que se ha buscado en los rasgos equivocados debido a que los estructurales son perceptuales, y generalmente forman parte de un procedimiento de identificación pero no del núcleo de muchos conceptos.

Finalmente, Fellbaum (1998) señala que un problema serio en el lexicón WordNet es que puede haber representados por el apuntador de hiponimia no una sino varias relaciones semánticas. Por ejemplo, Wierzbicka (1984)

distingue 5 tipos de hiponimia, donde dos de ellas son las más importantes: las relaciones representadas por las expresiones lingüísticas IS-KIND-OF y IS-USED-AS-A-KIND-OF. La primera de ellas, IS-KIND-OF, está relacionada con la taxonimia e imprime un orden jerárquico en un conjunto de conceptos. Por otro lado, IS-USED-AS-A-KIND-OF hace énfasis en aspectos funcionales cuando se consideran artefactos. Wierzbicka los denomina como taxonómico y funcional, respectivamente. Por su parte, Pustejovsky, como se comentó en líneas anteriores, los llama formal y télico. Un ejemplo de ambos tipos de relaciones coexistiendo en una definición es el siguiente:

> Un atizador (formalmente) es una varilla de metal que (funcionalmente) se usa para atizar brazas ardientes.

En ocasiones, el hiperónimo es formal, por ejemplo, *un canario es un pájaro*, y otras veces es télico, como en el caso de *un adorno es una decoración.* La pregunta que surge de la situación anterior es: ¿qué pasa si ambos tipos de hiperonimia están disponibles para un nombre en un esquema de representación de conocimiento? De momento, lo que podemos derivar de esta revisión final de capítulo es una posibilidad adicional de encontrar más de un hiperónimo asignado a un concepto, aunado a variantes léxicas, diferentes niveles de una estructura jerárquica de categorías y casos no consensuados.

Capítulo 4

Extracción de relaciones de hiponimia-hiperonimia

Este capítulo está enfocado a presentar detalles de la metodología que se propone para la extracción de un subconjunto de relaciones de hiponimia-hiperonimia a partir de textos de especialidad. Esta metodología incluye etapas que van desde el pre-procesamiento de la colección de textos de entrada hasta la obtención de relaciones léxico-semánticas de hiponimia-hiperonimia a partir de los hiperónimos más frecuentes.

El proceso para la extracción de relaciones de hiponimia-hiperonimia se divide básicamente en 4 etapas. A grandes rasgos, la primera etapa consiste en un preprocesamiento de la colección de textos de entrada que va desde la eliminación de algunos elementos de la fuente textual (imágenes, tablas, información en paréntesis, etc.) hasta correcciones posteriores al etiquetado de partes de la oración que permitan la obtención de mejores resultados. Posteriormente, se lleva a cabo una segunda etapa que combina un análisis sintáctico superficial, seguido por un filtro semántico con la finalidad de recuperar los mejores candidatos a CDs del corpus etiquetado. Una tercera etapa se enfoca en extraer los términos e hiperónimos de los fragmentos candidatos. Finalmente, una cuarta etapa utiliza los hiperónimos más frecuentes para realizar una fase de *bootstrapping* y con ello extraer categorías subordinadas al *genus*.

4.1 Características de la colección de textos de entrada

Una de las características más importantes de la fuente de información textual de entrada es que debe estar restringida a un dominio específico de conocimiento. Un dominio circunscribe su conocimiento a un conjunto de conceptos, que a su vez están representados lingüísticamente por unidades léxicas (términos). El establecimiento de relaciones entre los términos de un dominio contribuye a estructurar su espacio conceptual y esto a su vez da

cuenta del conocimiento acotado dentro de un dominio.

Aunado a lo anterior, surge la pregunta: ¿qué tipo de fuentes textuales son las más apropiadas?, ¿qué criterios se consideran para seleccionarlas? En este punto, Buitelaar y Cimiano (2008) señalan que no existe un consenso aún sobre el tipo de fuentes textuales que deban usarse como evidencia para soportar la construcción de recursos tales como tesauros u ontologías, sin embargo, en este trabajo consideramos que lo más recomendable es contar con documentos donde exista una probabilidad alta de que los conceptos que se manejan dentro del mismo sean explicitados, por ejemplo, por medio de una definición. De acuerdo con Cabré *et al.* (2000), el grado de especialización de un texto varía en un continuo desde un nivel bajo (prensa escrita y revistas de gran difusión), seguido por un nivel medio (artículos de divulgación científica, secciones técnicas de diarios y revistas), hasta un nivel alto (artículos científicos). En el mismo orden de ideas, Alarcón (2009) menciona que los textos especializados son una buena fuente para encontrar descripciones o explicaciones de conceptos de acuerdo con el tipo de situación comunicativa a la cual pertenecen. Por ejemplo, en la comunicación entre expertos y (semi)expertos, o bien entre expertos y no expertos se utilizan con mayor frecuencia contextos donde se aporta información sobre los atributos y las relaciones entre los términos. Dado lo anterior, asumimos que el grado de especialización más adecuado para la extracción de conocimiento debe ser de medio a alto, priorizando el nivel medio, porque es donde existe una probabilidad mayor de encontrar definiciones que reflejen, en mayor o menor grado, un consenso respecto a un concepto.

4.2 Preprocesamiento de la información de entrada

El preprocesamiento es una parte esencial para cualquier sistema de procesamiento de lenguaje natural debido a que los caracteres, palabras y oraciones que se identifiquen en esta etapa son las unidades fundamentales de las etapas posteriores, que van desde el análisis y etiquetado de componentes, hasta aplicaciones como la extracción de información y los sistemas de traducción máquina.

En los siguientes apartados se detallan los pasos de preprocesamiento considerados para la fuente de información de entrada.

4.2.1 Eliminación de algunos elementos de la fuente textual

Este paso corresponde a la eliminación manual de elementos de la colección de textos que por el momento no se consideran importantes. Entre estos se incluyen: imágenes, tablas, direcciones de correo electrónico, etc. La eliminación de tablas e imágenes se realiza de forma manual, sea al momento de seleccionar la información de algún repositorio (por ejemplo, de la Web) o después de haber recolectado el total de documentos. Para la eliminación de hipervínculos y correos electrónicos se aplican expresiones regulares. Por ejemplo, para eliminar direcciones de correo electrónico se utilizan expresiones como la siguiente:

```
text = re.sub(".*@.*"," ",text)
```

4.2.2 Delimitación de palabras

Las palabras se delimitan por un espacio. Por medio de un procedimiento semi-automático, los puntos correspondientes a abreviaturas se dejan ligados a ellas y en el caso del punto delimitando el límite de una oración se considera un espacio entre la palabra que lo precede y la que se encuentra después del punto. Del mismo modo, para todos los signos y marcas de puntuación (, ; : ¿ ? ¡ ! … () [] {}) automáticamente se asigna un espacio, de tal suerte que el etiquetador los considere como un *token* independiente.

4.2.3 Segmentación de oración

El proceso que se aplica a la fuente de información de entrada básicamente se relaciona con un proceso de *segmentación de oración* para determinar estas unidades de procesamiento. Esta tarea incluye la identificación de límites de oración. Debido a que la mayoría de los lenguajes escritos tienen marcas de puntuación que ocurren en los límites de una oración, la segmentación de oración se refiere comúnmente como *detección del límite de una oración, desambiguación del límite de oración* o *reconocimiento del límite de una oración.*

El tokenizador de oraciones que se utilizó para esta etapa es el *Punkt* que se encuentra disponible en el módulo NLTK. Este módulo puede ser entrenado con corpus en español, disponibles también en NLTK.

```
sent_tokenizer = nltk.data.load("tokenizers/punkt/spanish.pickle")
oraciones = sent_tokenizer.tokenize(corpus)
```

4.2.4 Eliminación de paréntesis con información complementaria

Con la meta de aclarar la terminología usada con frecuencia se detalla esta información mediante el uso de paréntesis incrustados en el contexto de una oración. La información entre paréntesis está relacionada generalmente con aclaraciones, acrónimos, raíces etimológicas, sinónimos e incluso definiciones breves. Los siguientes ejemplos ilustran las situaciones mencionadas con anterioridad:

1. *Electronigrafia* (ERG) es una prueba que mide la actividad eléctrica de la retina a la luz.
2. *Conjuntivitis* es la hinchazón (inflamación) o infección de la membrana que cubre los párpados.
3. El *dolor ocular* (que no se debe a una lesión) se puede describir como una sensación urente, pulsátil, dolorosa o lacerante ubicada en o alrededor del ojo.

Aunque esta información es valiosa, no se considera en este trabajo debido a que requeriría de expresiones regulares más complejas para capturar estos patrones en la gramática para la fase de análisis sintáctico superficial porque tiene como límite, por lo menos para la versión 2.0 del módulo NLTK, 100 etiquetas gramaticales. Por tanto, removemos automáticamente esta información del texto previo a la fase de etiquetado de partes de la oración.

4.2.5 Etiquetado de partes de la oración

El etiquetado de partes de la oración es un proceso que asigna una categoría gramatical o parte del discurso a cada palabra en un corpus. Por ejemplo, una salida típica de un etiquetador proporciona tres elementos: la palabra original (forma de palabra), la etiqueta gramatical y el lema, como lo muestra la tabla 8.

Tabla 8. Salida del etiquetador TreeTagger.

Forma de palabra	Etiqueta gramatical	Lema
Define	VLfin	Definir

El siguiente ejemplo muestra una oración etiquetada con el etiquetador TreeTagger (Schmid, 1994):

El/ART síntoma/NC característico/ADJ de/PDE rojez/NC ocular/ADJ ./FS

El etiquetador que proponemos para esta fase de etiquetado de partes de la oración es precisamente TreeTagger, dada su disponibilidad y precisión respecto a otros también disponibles para el Español. Barrón (2006) reporta un análisis del desempeño de dos etiquetadores disponibles gratuitamente: TreeTagger y Freeling, donde concluye que TreeTagger logra un mejor desempeño.

Problemas con el etiquetado de partes de la oración en dominios especializados

Uno de los problemas que se ha documentado en la literatura es la falta de precisión en el etiquetado de partes de la oración para dominios especializados. Dado que los corpus que generalmente se usan para entrenar estos etiquetadores son de lengua general, los niveles de precisión bajan significativamente en corpus especializados (Amrani *et al.*, 2004; Miller *et al.*, 2007). Un comportamiento semejante se tiene también cuando se desarrollan y entrenan etiquetadores para un dominio específico y luego se usan para etiquetar otro diferente.

Dado lo anterior, si bien no incluimos una etapa de entrenamiento debido a que no contamos con un corpus anotado manualmente, consideramos conveniente analizar los datos de una muestra aleatoria de 200 oraciones de un conjunto de documentos de medicina y el porcentaje de error obtenido al etiquetar por medio de TreeTagger. Sin duda, un alto porcentaje de etiquetado erróneo repercutiría negativamente en nuestros resultados, es por ello que resulta indispensable cuantificarlo.

Análisis de resultados de etiquetado

En términos generales, de las 200 oraciones analizadas, se obtuvo un total de

76% de oraciones correctamente etiquetadas. El 24% restante presentó por lo menos algún tipo de error de etiquetado.

Figura 15. Precisión del etiquetador TreeTagger.

La tabla 9 muestra los tipos específicos encontrados, así como su frecuencia absoluta y porcentual de ocurrencia. Es importante aclarar que se puede encontrar más de un error en una oración, por lo que la frecuencia total de 54 incluye todos los errores encontrados en el conjunto de oraciones con problemas de etiquetado.

Tabla 9. Errores de etiquetado con TreeTagger.

Etiqueta[16]		**Frecuencia absoluta**	**Porcentaje**
Incorrecta	**Correcta**		
NC	ADJ	16	29.6
VLFIN	NC	12	22.2
NP	NC	10	18.5
VLFIN	ADJ	4	7.4
ADJ	NC	3	5.6
NC	NP	2	3.7
ADV	ADJ	2	3.7
VLFIN	NP	1	1.9
ADJ	NP	1	1.9
VSFIN	NC	1	1.9
VLINF	NC	1	1.9
VLINF	ADJ	1	1.9
	Total:	54	100

A partir de la tabla 9 se puede observar que los errores más comunes son aquellos donde se etiqueta erróneamente una palabra como nombre común (NC) y la etiqueta correcta es un adjetivo (ADJ), así como también aquellos donde se etiqueta como un verbo (VLFIN) y la correcta es un nombre

[16] Etiquetas gramaticales de TreeTagger disponibles en la tabla 36 del apéndice.

común y, finalmente, las palabras que se etiquetan como nombre propio (NP) y en realidad son nombres comunes (NC). En conjunto, estos tres tipos de error conforman el 70.3%. Todos los errores están relacionados con palabras estrechamente asociadas con el dominio, por lo que evidencian la situación planteada por Amrani *et al.* (2004) y Miller *et al.* (2007).

Errores de lematización

La lematización es el proceso mediante el cual se agrupan diferentes formas flexionadas de una palabra de modo que se puedan analizar como un elemento único. Para los fines de este trabajo, un proceso de lematización preciso nos permitirá la obtención de mejores resultados a partir de las primeras etapas del análisis. Por ejemplo, podemos tener casos donde el hiperónimo *enfermedad* y *enfermedades* se cuenten como dos hiperónimos diferentes, de manera que si un porcentaje alto de los potenciales hiperónimos no se lematizan, puede sesgar la distribución de frecuencias de hiperónimos, útil en el tercer proceso de filtrado de CDs y como materia prima para extraer más hipónimos. Por otro lado, si no existe una lematización aceptable de adjetivos y nombres en la búsqueda de modificadores adjetivos y nominales que junto con el hiperónimo den lugar a hipónimos, las frecuencias tenderán a dispersarse en las flexiones de una palabra impactando negativamente en los resultados.

A partir de las oraciones de la muestra analizada se obtuvo que las categorías gramaticales que presentan más errores de lematización son precisamente el nombre común y los adjetivos. Del conjunto de 200 oraciones analizadas, 43 oraciones (22%) presentaron por lo menos un error de lematización. El error más común es que se incluye una gran cantidad de flexiones de género y número de nombres y adjetivos relacionados con el dominio. Por ejemplo, para el caso de la palabra *aeróbico*, después de etiquetar y lematizar con TreeTagger se tienen todavía los siguientes elementos: *aeróbico, aeróbica, aérobicos, aeróbicas*. Afortunadamente, estas flexiones son en su mayoría regulares, lo cual permite la aplicación de heurísticas para mejorar el proceso.

El siguiente apartado presenta las heurísticas propuestas para mejorar la lematización de nombres y adjetivos con flexiones regulares.

Heurísticas para mejorar el proceso de lematización

La tabla 10 muestra las heurísticas propuestas para mejorar el proceso de lematización de TreeTagger. El procedimiento que se describe en este apartado se basa en la información que se puede encontrar en el corpus, es decir, si se encuentra el nombre o adjetivo en singular, el proceso de lematización se hará más eficiente. Por otro lado, en las heurísticas sólo se consideran las categorías gramaticales de nombre común, nombre propio y adjetivo porque son las más importantes para nuestro análisis. En el caso de los verbos considerados en definiciones analíticas, que también son muy importantes, no se presentaron problemas de etiquetado o de lematización. Del mismo modo, otras categorías gramaticales tales como preposiciones, artículos, adverbios, etc., presentaron un porcentaje de error muy bajo. Finalmente, se restringe la longitud de la palabra a aquellas que sean mayores que 5 porque asumimos que a medida que la longitud de una palabra es mayor, si coincide con otra, mayor será la probabilidad de que se trate de la misma palabra pero flexionada.

Tabla 10. Heurísticas para mejorar lematización.

	Validaciones en código Python	
	If word1 != word2: #(aplica heurísticas 1-6)	**Ejemplo**
1	if len(word1)==len(word2) and word1[:-1] == word2[:-1] and word1[-1:] in vocal and word2[-1:] in vocal and word1[-1:] != word2[-1:]:	Cutáneo-cutánea (cutáne-)
2	if len(word1)==len(word1) and word1[:-2]== word2[:-2] and word1[-2:] in suffixPlural and word2[-2:] in suffixPlural and word1[-2:]!= word2[-2:]:	Borrosos-borrosas (borros-)
3	if len(word1)!=len(word2) and word1[:-1] == word2 and word2[-1:] in vocal and word1[-2:] in suffixPlural	Síndromes-sindrome (sindrome)
4	if word1[:-2] == word2[:-1] and word2[-1:] in vocal and word1[-2:] in suffixPlural:	Cefaleas-cefalea (cefalea))
5	if word1[:-2] == word2 and word2[-1:] in consonante and word1[-2:] in suffixPlural:	Natural-naturales (natural)
6	if word1[:-1] == word2 and word1[-1:] in vocal and word2[-1:] in consonante:	Vicerrectora-vicerrector (vicerrector)
7	Elif word1 == word2 and tagWord1 != tagWord2: #(Aplica solo en casos donde las etiquetas sean diferentes y correspondan a NC o NP).	Enfermedad/nc-Enfermedad/np (Enfermedad/nc)

Resultados de aplicación de heurísticas de lematización

Los resultados de la aplicación de las heurísticas detalladas en el apartado anterior dan como resultado una mejora del 82% en la lematización de las flexiones regulares de nombres y adjetivos. Estos datos se corroboraron con la muestra de 200 oraciones. Los casos donde no se lematiza son aquellos donde la longitud de palabra es menor o igual que 5 (por ejemplo, *renal*); palabras cuya forma singular no se encuentra en la colección de textos y en los casos de otras etiquetas gramaticales. La figura 16 muestra los resultados de lematización obtenidos. Para el caso de la muestra analizada se obtuvo un total de 42 oraciones donde por lo menos había una palabra no lematizada. En conjunto, las 42 oraciones generaron un total de 55 problemas de lematización. Finalmente, aplicando las heurísticas anteriores se logró una corrección del 82%, lo que ofrece un panorama más alentador para los procesos subsecuentes.

Figura 16. Resultados de aplicación de heurísticas para lematización.

Corrección de errores de etiquetado

Los errores de duplicación de etiquetas para una palabra donde se considera como nombre propio (NP) o nombre común (NC) fueron normalizadas a nombre común, con el argumento de que los nombres comunes son más frecuentes. Por ejemplo, la palabra *antioxidantes* presenta tres etiquetas dentro del corpus, el siguiente ejemplo incluye también la frecuencia con la que aparece con cada una de las etiquetas:

Antioxidantes → nc/12, np/4, adj/3

En esta situación, la etiqueta que se asigna es la de nombre común por ser la más recurrente en corpus (Bird *et al.*, 2009). Por tanto, para este caso particular, la etiqueta NP de *antioxidantes* será sustituida por NC. Los casos donde aparece como un adjetivo no se modifican. Esta modificación se incluye al final de las heurísticas para corregir lematización mostradas en la tabla 10. Por otro lado, el caso más frecuente donde se etiqueta un adjetivo como nombre común puede afectar nuestros resultados si se encuentra precediendo otro nombre común, por lo que puede ser extraído como potencial hiperónimo, sin embargo, esperamos que la ocurrencia de estos casos sea mínima. Finalmente, los casos que se etiquetan como verbos y realmente son adjetivos o nombres valiosos para el dominio no alcanzan a ser cubiertos por nuestras expresiones regulares en las dos etapas de nuestra metodología. Dado lo anterior, se considera importante la aplicación de herramientas como las propuestas por (Amrani *et al.*, 2004; Miller *et al.*, 2007) para disminuir estos problemas. Desafortunadamente no se encontró implementación alguna disponible de estos métodos, por lo que consideramos es indispensable que se realicen trabajos encaminados a resolver este tipo de problemas.

4.2.6 Normalización de etiquetas gramaticales

Con el objetivo de reducir el alcance de las expresiones regulares sólo a los verbos característicos de definiciones analíticas, se aplica un paso de normalización de algunas etiquetas. La tabla 11 muestra los cambios a verbos y otros elementos que resultan también de interés diferenciar dentro de las definiciones.

Tabla 11. Normalización de etiquetas.

Palabra o frase modificada	**Etiqueta de TreeTagger**	**Nueva etiqueta**
Verbos usados en definiciones analíticas	Vlfin, Vlinf, Vlger, Vladj	Vlfind
De, de + art, Contracción de + art (del)	Prep Prep + art Pdel	Pdel
O (disyunción)	CC	Cco
A	Prep	Pa
.	Fs	Fsp

4.3 Fase de extracción conceptual

Los CDs, concretamente aquellos que siguen un patrón analítico, son la materia prima de la metodología que aquí se expone. En la sección 1.6.4 se presentó información importante relacionada con CDs, qué son, cuáles son sus elementos constituyentes, su estructura sintáctica, qué verbos son los más característicos, etc. Sierra *et al*. (2008) plantean que es posible la extracción automática de CDs a partir de considerar patrones verbales que se usan con frecuencia en la definición de conceptos. La herramienta Ecode representa un ejemplo de software enfocado a este tipo de tareas de extracción conceptual para el español. Entre los resultados mostrados por Ecode está una clasificación de CDs en analíticos, funcionales, extensionales y sinonímicos. Finalmente, los resultados generados por la herramienta hacen énfasis en delimitar el término y la definición, dependiendo del tipo de CD.

Una de las estrategias a seguir en nuestra metodología es la extracción de todas las oraciones delimitadas por un "." que contengan un verbo característico de definición analítica donde haya por lo menos una longitud de oración de 4 o mayor. La heurística anterior se fija a partir de considerar que el candidato a CD más corto (en número de palabras) debe tener un verbo característico de definición analítica, un término (el más simple es un nombre) y una FN (otro nombre) precedido por un artículo: *Perro es un animal*. Este punto de partida lo consideramos como nuestro *baseline*. A partir de este *baseline*, proponemos filtrar todas aquellas oraciones donde exista una estructura sintáctica específica, como se detalla en el siguiente apartado.

Nuestra propuesta para extraer CDs analíticos difiere del método de Alarcón (2009) en que utiliza *a priori* como filtro la estructura sintáctica de los elementos más importantes de una definición analítica: término, patrón verbal e hiperónimo. Además, consideramos también una tipificación sintáctica de otros elementos que se usan regularmente en definiciones analíticas, como son los sinónimos y los patrones pragmáticos. Esta estructura sintáctica se plasma en una gramática *chunk* que se enfoca en extraer los candidatos que cumplen con las restricciones impuestas. Aunado a lo anterior, para eliminar *ruido* en los resultados proponemos dos filtros más, uno semántico, que elimina núcleos nominales indicadores de otro tipo de relación (causal o parte-todo), y

otro más que utiliza los hiperónimos para filtrar los CDs más importantes.

4.3.1 Primer filtro de extracción conceptual

El enfoque para extraer el término y el hiperónimo de un CD analítico consiste en tomar en cuenta el patrón contextual del verbo para ubicar los elementos de interés. Para lograr lo anterior, se consideró un conjunto de expresiones regulares implementadas en una gramática *chunker*. Los elementos de estas expresiones regulares se modelaron vía considerar cada uno de ellos como un constituyente. Un constituyente es una palabra o grupo de palabras que funciona como una unidad dentro de una oración (Jurafsky y Martin, 2009). Entre los constituyentes que consideramos se encuentran los siguientes: el término, sinónimos del término, patrones pragmáticos y la FN donde se puede localizar el hiperónimo.

La decisión de diseñar la gramática de forma manual se basa principalmente en la disposición que se tiene actualmente de análisis precisos sobre la estructura sintáctica de definiciones analíticas (Aguilar, 2009), así como también de los diferentes patrones contextuales de los verbos más característicos de este tipo de definición, lo que permite identificar y extraer los elementos de interés. Además, se cuenta con una tipificación de la estructura más común de términos (Vivaldi, 2001) y de FNs donde se puede localizar el hiperónimo (Wilks *et al.*, 1996). Otro aspecto importante que es necesario mencionar es que los patrones verbales para introducir definiciones analíticas se consideran independientes de dominio, tal vez con variantes estilísticas, sin embargo, en esencia siguen un patrón regular. Por otro lado, el aprendizaje automático de la gramática de patrones requeriría de un análisis sintáctico más profundo para determinar los constituyentes implícitos en una oración, y actualmente estos analizadores se encuentran por debajo de un 90% de precisión (Ballesteros *et al.*, 2010).

Los siguientes apartados presentan una descripción general sobre la fase de análisis sintáctico superficial (shallow parsing) propuesto para la extracción de CDs, así como también de la colección de definiciones utilizada, que conjuntamente con los análisis sintácticos realizados por Aguilar (2009), se tomaron en cuenta para construir la gramática de la fase de análisis sintáctico superficial para el proceso de extracción. Finalmente, se presentan

cada uno de los constituyentes modelados vía etiquetas gramaticales.

Chunking

Chunking es el proceso de identificar y clasificar segmentos de una oración por medio de la agrupación de partes gramaticales que forman frases no recursivas básicas (Bird *et al.,* 2009). Las frases o reglas conforman una gramática *chunk.* Por ejemplo, la regla <ART>?<NC>+<ADJ>* es la estructura de una FN donde el artículo es opcional, por lo menos debe tener un nombre y cero o más adjetivos. Los patrones de etiquetas son similares a las expresiones regulares, donde los símbolos * significan cero o más ocurrencias, + representa por lo menos una ocurrencia y ? indica la opcionalidad de un elemento.

La tabla 12 muestra una oración con partes gramaticales agrupadas. Por ejemplo, una FN puede ser un nombre simple o una estructura más compleja: determinante + nombre + FP. Adicionalmente, el núcleo verbal de la frase verbal (FV) es el verbo *ser*.

Tabla 12. Chunking de una oración.

<table>
<tr><th>Conjuntivitis</th><th>es</th><th>una</th><th>inflamación</th><th>de la conjuntiva</th></tr>
<tr><td>(NC)
Nombre común</td><td><VSFIN>
Verbo</td><td><DET></td><td><NC></td><td><PDEL><ART><NC></td></tr>
<tr><td rowspan="3">FN</td><td rowspan="2">Núcleo de FV</td><td colspan="2"></td><td>FP</td></tr>
<tr><td colspan="3">FN</td></tr>
<tr><td colspan="4">FV</td></tr>
</table>

Construcción de la gramática chunk

En este trabajo, asumimos que la estructura sintáctica de los CDs es lo suficientemente regular, de tal suerte que es posible modelar los diferentes elementos de interés (término, patrón verbal e hiperónimo) mediante expresiones regulares que consideren etiquetas gramaticales. Para lograr lo anterior, se diseñó manualmente una gramática de expresiones regulares tomando en cuenta el análisis sintáctico de definiciones analíticas realizado por Aguilar (2009) y 1477 CDs extraídos de Wikipedia de áreas tales como biología, medicina, ingeniería y lingüística. Las definiciones de Wikipedia no siguen un patrón definicional formal comparado, por ejemplo, con un diccionario, y son más cercanas a las definiciones encontradas en dominios especializados. Además, Wikipedia es un recurso donde voluntarios de todo el mundo pueden escribir artículos respecto a una gran variedad de temas, lo que

garantiza, en parte, que el comportamiento más común de CDs va a estar presente.

El procedimiento seguido para extraer las definiciones de Wikipedia fue localizar el CD de un término del dominio y acceder a todos los hipervínculos de los conceptos que se encontraran en dicha página para extraer sus CDs verificando que se tratara de CDs analíticos. Este proceso se repitió en las páginas nuevas accedidas hasta obtener el conjunto de definiciones para construir la gramática.

Finalmente, con la gramática diseñada se logró capturar el 77% del comportamiento más canónico de los 1477 CDs. Es importante resaltar que un 99% de los términos definidos tiene la estructura mostrada en la tabla 13, lo que representa una evidencia empírica importante para tomar en cuenta sólo estos patrones para términos en nuestra metodología. Los CDs restantes (23%) básicamente presentan por lo menos un error de etiquetado en el fragmento donde se localiza el término y el hiperónimo, o alguna estructura de término no considerada.

Estructura de términos

En tareas de extracción terminológica para catalán (Estopà, 2003) y español (Vivaldi, 2001), los patrones más comunes para construir términos compuestos son aquellos que consideran modificadores adjetivos, nombres y de FP, específicamente éstos últimos con la preposición "de". La tabla 13 muestra los patrones considerados en esta investigación, así como también ejemplos de cada uno de ellos.

Tabla 13. Patrones terminológicos

Patrón	**Ejemplo**
Nombre + Adjetivo	Enfermedad cardiovascular
Nombre + FP	Enfermedad de Alzheimer
Nombre + nombre	Diabetes mellitus
Acrónimos	SIDA
Nombre + Letra	Vitamina A
Letra + nombre	H Pylori

La preposición "de" se incluye básicamente por dos razones principales: es la más recurrente, por lo menos en inglés (Jurafsky, 2009; Litkowski, 2002) y español, así como también la más usada para la construcción de términos en

español (Vivaldi, 2001).

Estructura de hiperónimos

De acuerdo con Wilks *et al.* (1996), las definiciones de nombres en diccionarios se escriben normalmente de tal forma que se puede identificar el *genus* o hiperónimo del término que se define. Por ejemplo, en la siguiente definición:

1. Conjuntivitis – Una inflamación de la conjuntiva del ojo…

Inflamación es el *genus* o hiperónimo del término *conjuntivitis* y el resto de la oración corresponde a la *differentia*. La heurística parece lógica e intuitiva, sin embargo, existe una buena cantidad de definiciones donde el núcleo de la FN no es un hiperónimo apropiado del término. Por ejemplo,

2. Piedra preciosa – Cualquiera de varios minerales…
3. Microscopio electrónico – tipo de microscopio que utiliza una partícula…

En las definiciones 2-3, *cualquier* y *tipo* son núcleos vacíos. Estos núcleos se deben filtrar con el objetivo de encontrar un hiperónimo que se relacione semánticamente con el término. En el caso de (2), el hiperónimo *mineral* es el más relacionado con el término definido. Por otro lado, para el caso (3), el núcleo *tipo* no es un hiperónimo, en lugar de ello, es un núcleo vacío indicador de un tipo de hiponimia (Croft y Cruse, 2004; Wierzbicka, 1996). En resumen, los hiperónimos en 2 y 3 se localizan después de la preposición "de". Así, esto justifica la consideración de FNs con modificador de FP con núcleo "de". La identificación de estos elementos no relevantes que preceden el hiperónimo se facilita con el etiquetado gramatical.

Por otro lado, cuando no existe un núcleo nominal y se encuentra un determinante como *aquel* seguido por un pronombre relativo como *que*, *cuyo*, el hiperónimo se extrae de la FN del término, es decir, el núcleo de esta frase será el hiperónimo. En el siguiente ejemplo no existe un nombre después del determinante *aquel*, por lo que el hiperónimo se extrae de la frase *tumor invasivo* y, en este caso, el hiperónimo corresponde a *tumor*.

Un tumor invasivo es aquel que se extiende a áreas circundantes…

Finalmente, una excepción la constituyen los casos donde la FN del término definido consta de un simple nombre. En estos casos no se realiza ninguna extracción.

Sinónimos

Es una práctica relativamente común en Wikipedia incluir por lo menos un sinónimo del término definido. Por ejemplo,

1. La andropausia o menopausia masculina, es el proceso por el cual las capacidades sexuales…
2. El vomito, también llamado emésis, es la expulsión violenta y espasmódica del contenido del estomago a través de la boca…
3. Se denomina labio leporino o fisura labial al defecto congénito, que consiste…

Dado este patrón de comportamiento en definiciones, se consideró importante su inclusión como un constituyente con cero o más ocurrencias dentro de una definición.

Patrón pragmático

En algunos casos, es común encontrar adverbios o frases compuestas que señalen el alcance de una definición o el área de conocimiento a la que pertenecen tanto el término como la definición, a estas expresiones se les denomina patrones pragmáticos (Alarcón, 2009). Los siguientes ejemplos muestran variantes de una misma definición donde el patrón pragmático tiene diferentes posiciones:

1. En Matemáticas, la Teoría Lineal puede ser considerada como una primera aproximación de una descripción teórica completa acerca del comportamiento del oleaje.
2. La Teoría Lineal en Matemáticas puede ser considerada como una primera aproximación de una descripción teórica completa acerca del comportamiento del oleaje.
3. La Teoría Lineal, en Matemáticas, puede ser considerada como una primera aproximación de una descripción teórica completa acerca del comportamiento del oleaje.

Los patrones pragmáticos que consideramos son los que se encuentran antes del término y los que están insertos entre el término y el patrón verbal porque son los más comunes en el conjunto de CDs.

Modelado de constituyentes

A continuación, se presentan las expresiones regulares correspondientes a cada uno de los elementos descritos en los apartados anteriores. Las etiquetas son las utilizadas por TreeTagger; sin embargo, la normalización y los cambios de etiquetas propuestos en la sección 4.2.6 ya se ven reflejados en estas expresiones regulares.

Término

$$Term: \langle art \rangle ? \langle alfs \rangle ? \langle nc \mid \text{np} \mid \text{acrnm} \mid \text{pe} \rangle + \langle adj \mid \text{alfs} \mid \text{card} \rangle *$$
$$\left(\langle pdel \rangle \langle nc \mid \text{np} \mid \text{adj} \mid \text{acrnm} \rangle + \right) *$$

Sinónimos

$$Sin1: \begin{pmatrix} \langle cm \rangle ? \langle cco \rangle \langle adv \rangle ? \langle art \rangle ? \langle nc \mid \text{np} \mid \text{adj} \mid \text{acrnm} \mid \text{alfs} \mid \text{pe} \rangle + \\ \left(\langle pdel \rangle \langle nc \mid \text{np} \mid \text{adj} \mid \text{acrnm} \rangle + \right) ? \end{pmatrix}$$

$$Sin2: \begin{pmatrix} \langle cm \rangle ? \langle cco \rangle ? \langle vladj \mid \text{vlfind} \mid \text{adv} \rangle + \langle csubx \rangle ? \langle art \rangle ? \\ \langle nc \mid \text{np} \mid \text{adj} \mid \text{acrnm} \mid \text{alfs} \mid \text{pe} \rangle + \\ \left(\langle pdel \rangle \langle nc \mid \text{np} \mid \text{adj} \mid \text{acrnm} \rangle + \right) ? \end{pmatrix}$$

Patrón pragmático

$$Ppragm: \left(\langle cm \rangle ? \langle pen \mid \text{prep} \rangle \langle art \rangle ? \langle nc \mid \text{np} \mid \text{adj} \mid \text{acrnm} \rangle + \left(\langle pdel \rangle \langle nc \mid \text{np} \mid \text{adj} \rangle + \right) ? \right)$$

FN del hiperónimo

$$Hiper1: \langle art \mid \text{qu} \mid \text{dm} \rangle ? \langle adj \rangle ? \langle nc \rangle + \langle adj \rangle * \left(\langle adv \mid \text{ccad} \rangle \langle adj \rangle \right) ?$$
$$\left(\langle pdel \rangle \langle nc \mid \text{adj} \rangle + \right) *$$

$$Hiper2: \langle dm \mid \text{art} \rangle \langle cque \mid \text{rel} \rangle$$

$$Hiper3: \langle art \mid \text{card} \rangle \left(\langle pdel \rangle \langle nc \mid \text{adj} \rangle \right) + \left(\langle adv \mid \text{ccad} \mid \text{adj} \rangle \right) ?$$

Expresiones regulares y patrones contextuales

De acuerdo con Alarcón (2009), para identificar automáticamente los

elementos constitutivos de un CD se debe tomar en cuenta que términos y definiciones tienden a seguir ciertos patrones de posición en relación con el verbo que los introduce. Por ejemplo, con el verbo *ser*, generalmente el término se encuentra antes del verbo y el hiperónimo después.

La tabla 14 muestra las expresiones regulares de los fragmentos de CDs que son de nuestro interés en el marco de un patrón contextual específico. Cada constituyente se representa por el nombre genérico asignado en la descripción de constituyentes proporcionada en los apartados anteriores. Por ejemplo, la etiqueta que representa al término es *Term*; el patrón pragmático es *Ppragm*; los sinónimos están representados por *Sin1*, *Sin2,* y las frases donde se puede encontrar el hiperónimo son *Hiper1*, *Hiper2 e Hiper3*. Las etiquetas *vsfin*, *vlfind*, y *vmfin* representan el verbo *ser*, verbos usados en definiciones analíticas (ver tabla 5, capítulo 1) y verbo modal, respectivamente. Las etiquetas restantes pueden ser consultadas en la tabla 36 del apéndice.

Tabla 14 . Expresiones regulares del fragmento de CD.

Verbo	**Expresión regular del fragmento de CD**	
Ser	①	<fsp><Term><Sin1|Sin2>*<Ppragm>?<cm>?<cque>?<vsfin><Hiper1|Hiper2|Hiper3>
	②	<fsp><Ppragm><Term><Sin1|Sin2>*<cm>?<cque>?<vsfin><Hiper1|Hiper2|Hiper3>
Definir, Considerar, referir, conocer, llamar, denominar, describir, Nombrar, Caracterizar, concebir, considerar, entender	③	<fsp><se><ppc>?<vlfind><csubx>?<Term><Sin1|Sin2>*<Ppragm>?<cm>?<pa|pal><Hiper1|Hiper2|Hiper3>
	④	<fsp><Term><Sin1|Sin2>*<Ppragm>?<cm>?<se><vlfind><pa|pal><Hiper1|Hiper2|Hiper3>
	⑤	<fsp><Term><Sin1|Sin2>*<Ppragm>?<cm>?<se>?<vsfin>?<vlfind><csubx><Hiper1|Hiper2|Hiper3>
	⑥	<fsp><Term><Sin1|Sin2>*<Ppragm>?<cm>?<vmfin><vsinf><vlfind><csubx><Hiper1|Hiper2|Hiper3>
	⑦	<fsp><Ppragm><se><ppc>?<vlfind><Term><Sin1|Sin2>*<pa|pal><Hiper1|Hiper2|Hiper3>
	⑧	<Sujeto><vlfind><Term><csubx><Hiper1|Hiper2|Hiper3>
	⑨	<fsp|cm><Hiper1><se><vlfind><csubx>?<Term>
	⑩	<fsp><Term><Sin1|Sin2>*<fsp|semicolon><vsfin><Hiper1|Hiper3>

Por limitaciones técnicas[17], las expresiones regulares 1 hasta la 7 que contemplan las estructuras de hiperónimos (Hiper1, Hiper2 e Hiper3) se manejaron realmente como tres reglas separadas cada una y comprenden los patrones más canónicos con el verbo *ser* y verbos en predicación secundaria. De esta manera, la gramática consta de 27 reglas. Por ejemplo, en el caso de la expresión regular 1, el fragmento debe estar precedido por un punto "." (<fsp>), como se trata del verbo *ser* (<vsfin>) el término debe estar ubicado en posición preverbal, con la posibilidad de cero o más sinónimos del término (<Sin1|Sin2>) y un patrón pragmático opcional (<Ppragm>); de igual forma, una coma "," (<cm>) o una partícula que (<cque>) pueden preceder al verbo; finalmente, se consideran tres posibles estructuras de frase para el hiperónimo (<Hiper1|Hiper2|Hiper3>), con lo que se tienen tres reglas. Un ejemplo de este tipo de regla es el siguiente:

> La brucelosis, también llamada fiebre malta o fiebre ondulante, es una enfermedad que ataca a muchas especies…

Por otro lado, la expresión regular 8, que conforma tres reglas, representa el patrón más canónico que tiene un sujeto en posición preverbal más un objeto y su predicado, ambos después del verbo. Un ejemplo del tipo de casos que cubre este patrón es el siguiente:

> El Diccionario médico de Stedman define la psicosis como un desorden mental severo, con o sin un daño orgánico…

La expresión regular 9 cubre los casos donde el hiperónimo se encuentra en una posición preverbal y el hipónimo se localiza después del verbo. Por ejemplo:

> Un problema hepático llamado colestasis puede ocasionar problemas con los niveles de vitamina D.

Finalmente, la expresión regular 10 corresponde a un patrón con el verbo *ser* que contempla un término separado de su definición mediante un signo de puntuación, que puede ser un punto "." o dos puntos ":".

[17] Como se mencionó en la sección 4.2, la versión de NLTK 2.0 considera un límite de 100 etiquetas.

Acrorigosis: es la sensación banal permanente y simétrica de manos y pies fríos.

En términos prácticos, para realizar análisis sintácticos superficiales, herramientas como NLTK permiten asignar a cada regla una etiqueta y ésta puede ser de utilidad para determinar el patrón verbal presente en el fragmento del CD candidato y con ello proceder a su partición para extraer los elementos más relevantes: término e hiperónimo.

4.3.2 Segundo filtro de extracción conceptual

Como se mencionó en secciones anteriores, los verbos que se proponen como característicos de definiciones analíticas no solo se usan para enunciar definiciones, el ejemplo más ilustrativo es el verbo *ser*. Dada esta situación, se propuso el primer filtro sintáctico que tiene como meta garantizar la existencia de un término (con una estructura sintáctica específica), un patrón verbal y un hiperónimo (implícito también en una estructura sintáctica específica). Sin embargo, a nivel semántico, es posible encontrar casos donde los elementos involucrados (término y potencial hiperónimo) no proyecten una relación de hiponimia-hiperonimia. Por ejemplo: *la cistinosis es la causa más común del síndrome de Fanconi.* En casos como el anterior tenemos una relación causal indicada por el nombre "causa", que, de acuerdo con Girju (2003) da cuenta de una causalidad explícita. Por otro lado, otro tipo de relación que puede ser predicada con el verbo *ser* es la de meronimia-holonimia: *El sistema nervioso simpático es parte del sistema nervioso autónomo.* En consecuencia, para enfrentar estos problemas se propone un segundo filtro de núcleos nominales indicadores de relaciones de meronimia-holonimia y causalidad, que son dos de las relaciones más explotadas en tareas de extracción automática y que se predican comúnmente con el verbo *ser*. En los siguientes apartados se describe el procedimiento seguido para obtener estos núcleos nominales.

Núcleos nominales indicadores de meronimia-holonimia

Si el núcleo nominal de la FN es un nombre como: *parte*, *componente*, *segmento, porción* y *fracción*, dichos elementos sugieren una relación de meronimia-holonimia donde el holónimo se puede encontrar después de la preposición "de": *El hígado es una parte del cuerpo*.

Con la finalidad de obtener evidencia empírica sobre la relación e importancia de estos núcleos nominales, así como también para generar una lista más amplia de elementos, se utilizó la herramienta *Sketch Engine*[18].

Mediante la herramienta *Thesaurus* y corpus en español, se generó un tesauro distribucional para las siguientes 5 palabras semilla: *parte*, *componente, segmento, porción* y *fracción*. Con los 5 tesauros anteriores se obtuvieron nuevos elementos: *pieza, trozo, fragmento*, *pedazo, capa, superficie,* etc. La tabla 15 resume la información obtenida donde sólo se muestran los nombres indicadores de meronimia-holonimia que aparecen en cada tesauro distribucional.

Es importante aclarar que las palabras *parte* y *miembro* no aparecen en la tabla 15 debido a que en su tesauro distribucional no ocurre ninguna de las palabras consideradas, salvo en el caso de *miembro* donde sí aparece el elemento *parte*. Sin embargo, en la literatura sobre relaciones de meronimia-holonimia se ha documentado que *parte* está estrechamente asociado con este tipo de relación (Cruse, 1986; Winston *et al.*, 1987; Wierzbicka, 1996; Girju y Badulescu, 2006). Además, de acuerdo con Wierzbicka (1996), *parte* se considera como un primitivo semántico representativo de relaciones parte-todo. Todo lo anterior se encuentra soportado también por la evidencia empírica presente en los CDs analizados para construir la gramática *chunk*.

En conclusión, se puede observar que los núcleos nominales tienen una alta similaridad de contextos gramaticales y colocacionales, por lo que estos resultados nos dan la pauta para su utilización en una segunda etapa de filtrado.

[18] Este recurso puede ser consultado en http://www.sketchengine.co.uk/

Tabla 15. Palabras relacionadas con nombres indicadores de meronimia.

Palabra	Palabras relacionadas en tesauro distribucional
Componente	Pieza, parte
Segmento	Componente, porción, parte
Porción	Trozo, segmento, fragmento, pedazo, parte, pieza, fracción, componente
Fracción	Porción, componente, capa, superficie
Fragmento	Trozo, pedazo, porción, pieza, parte, componente
Capa	Superficie, pieza, componente, parte, trozo
Pedazo	Trozo, fragmento, porción, pieza, parte, capa
Trozo	Pedazo, fragmento, porción, pieza, capa, superficie
Pieza	Trozo, fragmento, porción, parte, capa
Superficie	Parte, pieza, capa, componente
Bola	Trozo, pieza, capa

Núcleos nominales indicadores de relaciones de causalidad.

De acuerdo con Girju (2003), las construcciones causales pueden ser explícitas o implícitas. Generalmente, los patrones de causalidad explícita pueden contener palabras clave relevantes tales como *causa, efecto, consecuencia*, pero también ambiguos como *generar, inducir,* etc. En los CDs con estructura sintáctica analítica se pueden localizar nombres en la FN del término y también del hiperónimo que indican este tipo de relación: *causa, consecuencia, efecto,* etc.:

1. La cistinosis es la causa más común del síndrome de Fanconi.
2. La consecuencia del aumento de la presión intraventricular es el establecimiento del desnivel de presión…

Dado la situación anterior, para generar una lista de nombres indicadores de relaciones de causalidad se aplicó el mismo procedimiento mencionado en el apartado anterior. Las palabras *semilla* en este caso son: *causa*, *consecuencia* y *efecto*. La tabla 16 muestra el resumen de resultados.

Tabla 16. Núcleos nominales indicadores de causa-efecto.

Palabra	Palabras relacionadas
Causa	Consecuencia, razón, efecto, resultado, respuesta
Consecuencia	Causa, efecto, resultado
Efecto	Resultado, consecuencia, producto, causa, respuesta
Resultado	Efecto, respuesta, producto
Respuesta	Resultado
Causante	Causa, consecuencia, efecto, signo
Signo	Causa, consecuencia, resultado, efecto
Repercusión	Consecuencia, efecto, resultado
Secuela	Consecuencia, causa, efecto, resultado
Fruto	Producto, resultado, consecuencia
Síntoma	Signo, causa, consecuencia, efecto, señal, indicio, resultado, respuesta
Señal	Signo, respuesta, resultado, efecto, producto
Indicio	Signo, síntoma, consecuencia, antecedente, causa, motivo, señal, razón, resultado, causante
Antecedente	Causa, motivo, consecuencia, resultado, signo, síntoma
Razón	Motivo, causa, consecuencia, resultado, respuesta
Motivo	Razón, causa, consecuencia, resultado, efecto

En resumen, dado lo mencionado en los apartados anteriores, proponemos un filtro donde estos núcleos nominales se identifiquen de la lista de CDs resultado de la fase de análisis sintáctico superficial, ya sea que se encuentren en la FN del término o en la frase del hiperónimo. Si alguno de ellos está presente, entonces el CD se elimina de este conjunto.

4.4 Fase de extracción de hiperónimos

La etapa de extracción de hiperónimos se basa en el filtrado de núcleos vacíos o confusos. Cuando se localiza un núcleo vacío se procede a la búsqueda del hiperónimo más relacionado semánticamente. Si en esta fase no se extrae un hiperónimo candidato, el CD se elimina del conjunto, por lo que también es útil esta etapa para filtrar CDs no relevantes. A continuación se presenta una descripción de este tipo de núcleos nominales.

Núcleos nominales indicadores de hiponimia-hiperonimia

El núcleo nominal de la FN donde se puede localizar el hiperónimo puede ser un nombre que no corresponda precisamente a un hiperónimo, tal es el caso de *tipo*, *clase*, etc., cuando se encuentran precediendo una FP con núcleo “de”. Por ejemplo, *la aspirina es un tipo de medicamento denominado salicilato*. En

estos casos, como se menciona en Wilks *et al.* (1996), el hiperónimo debe buscarse después de la preposición "de". Para el caso del ejemplo anterior, el hiperónimo corresponde a *medicamento*. Los núcleos nominales que se consideran como indicadores de hiponimia son *tipo, forma, clase, especie, prototipo, variedad* y sus derivados: *subtipo, subclase, subespecie*, etc.

Por otro lado, núcleos colectivos como *conjunto*, *serie, colección, familia* y *grupo* si no van acompañados de una lista de FNs después de la preposición "de", el nombre después de la FP puede considerarse como un hiperónimo, que es el caso del siguiente ejemplo:

> La leucemia o leucosis es un grupo de enfermedades malignas de la médula ósea…

En Wilks *et al.* (1996) se presenta una discusión respecto a si son realmente núcleos vacíos o simplemente *confusos* en el sentido de que si encontramos una lista de elementos después de la preposición "de", entonces, en lugar de hablar de una relación IS-A, tenemos una relación parte-todo donde se deben crear ligas etiquetadas como TIENE_MIEMBRO de cada elemento enumerado en la diferencia específica al concepto definido, como sería el caso del siguiente ejemplo:

Una vajilla es un conjunto de platos, vasos, y tazas…

En nuestro caso particular, sólo extraemos el nombre después de la preposición "de" debido a que asumimos que en un contexto sintáctico de definición analítica existe una baja probabilidad de encontrar una lista de elementos y que esto nos proporcione otro tipo de definición (parte-todo). Para soportar lo anterior, obtuvimos datos del corpus esTenTen disponible en *Sketch Engine*. *Sketch Engine* tiene disponibles corpus enormes en varios lenguajes (inglés, español, francés, alemán, etc.) para realizar diferentes tipos de análisis lingüísticos o de procesamiento de lenguaje. Para efectos de esta investigación, se consideró un corpus en español con un tamaño de 2 mil millones de palabras (esTenTen). La expresión regular utilizada para la extracción de concordancias garantiza que haya una FN con un artículo y por lo menos un nombre obligatorio, así como también de cero hasta dos adjetivos, seguido del verbo *ser* más un artículo obligatorio y el nombre colectivo (conjunto, serie, familia, grupo,colección) precediendo la preposición "de":

[tag="ART"][tag="NC"]{1,2}[tag="ADJ"]{0,2}"es" [tag="ART"] [colectivo] “de”

Del conjunto de resultados se obtuvo una muestra aleatoria de 250 concordancias. En algunos casos el número total obtenido fue inferior a este umbral establecido de forma arbitraria. El resultado del análisis se presenta en la tabla 17. De la tabla es claro que la confiabilidad mayor es para los nombres familia (100%) y colección (100%), seguidos por grupo (97%) y *serie* (94%). Sin embargo, el nombre *conjunto* presenta también una alta confiabilidad como proveedora de relaciones de hiponimia-hiperonimia (91%). Los porcentajes están calculados con relación al número total de concordancias relevantes, es decir, aquellas donde se proporciona información importante sobre algún concepto.

Tabla 17. Análisis de concordancias.

Núcleo nominal	**Total de concordancias**	**Concordancias relevantes**	**Hiperónimo**	**Meronimia-Holonimia**
Conjunto	250	195	91%	9%
Serie	214	62	94%	6%
Grupo	237	130	97%	3%
Familia	17	6	100%	0%
Colección	142	45	100%	0%

Caso de un determinante más relativo sustituyendo al hiperónimo

En los casos donde no existe un núcleo nominal y se encuentra un determinante como *aquel* seguido por un pronombre relativo como *que*, *cuyo*, se considera que el determinante se encuentra focalizando al núcleo nominal del término definido, por lo que el hiperónimo se extrae de la FN del término, es decir, el núcleo de esta frase será el hiperónimo. En el siguiente ejemplo no existe un nombre después del determinante *aquel*, por lo que el hiperónimo se extrae de la frase *tumor invasivo* y, en este caso, el hiperónimo corresponde a *tumor*.

Un tumor invasivo es aquel que se extiende a áreas circundantes…

Finalmente, una excepción la constituyen los casos donde la FN del término definido consta de un simple nombre. En estos casos no se realiza ninguna extracción.

Otros núcleos vacíos

Existen otros núcleos nominales que pueden considerarse como *vacíos* y que indican la introducción posterior de un potencial hiperónimo en una estructura sintáctica de definición analítica, entre éstos se encuentran: *nombre* y *ejemplo*.

- Virus Ébola es el nombre de un virus de la familia Filoviridae y género Filovirus…
- El Paracetamol es un ejemplo de acetaminofeno, un tipo de medicación que ha demostrado ser un buen calmante del dolor.

Otro nombre que se encuentra con frecuencia en este tipo de estructuras sintácticas es *término*, sin embargo, en la mayoría de las ocasiones no introduce un hiperónimo sino los contextos de uso del término definido, por ejemplo:

- El prolapso de la válvula mitral es el término usado cuando la válvula no se cierra adecuadamente y puede ser causado por muchas cosas diferentes.

Hiperónimos como filtro de candidatos a CDs

Como parte de esta metodología proponemos el uso de los hiperónimos más frecuentes como filtro para obtener los candidatos a CDs más relevantes. Para lograr lo anterior, se pueden establecer umbrales de frecuencia de aparición del hiperónimo en la distribución de frecuencias de hiperónimos obtenidos de CDs. El supuesto que se considera aquí es que los hiperónimos que más ocurren en el conjunto de CDs son los que tienen mayor probabilidad de ser realmente hiperónimos del dominio. Así, estos hiperónimos más frecuentes pueden ser de utilidad para filtrar CDs no relevantes y aumentar la precisión de la extracción (Acosta *et al.*, 2011).

Asumimos que las colecciones de textos que se analicen estarán relacionadas con una temática específica, además, mientras mayor sea el

tamaño de la colección de textos, mejores resultados podrán obtenerse con la aplicación de este tipo de metodologías.

4.5 Extracción de hipónimos a partir de hiperónimos

El *genus* o hiperónimo es una categoría genérica que clasifica los términos de un dominio. Es común encontrar un término descrito en función de una categoría más específica derivada a partir de vincular el hiperónimo con un modificador, sea éste un nombre, un adjetivo o una FP. Los siguientes dos ejemplos reflejan esta situación:

1. La deuteranopia o deuteranopsia es una **disfunción visual** consistente en alteración para la percepción del color…
2. La deuteranopia o deuteranopsia es una **disfunción de la vista** consistente en alteración para la percepción del color…

Las definiciones 1 y 2 son sinónimas porque las FNs del hiperónimo son variantes léxicas o variaciones de permutación (Vivaldi, 2001) del mismo concepto en el sentido de que el adjetivo relacional *visual* tiene su origen en el nombre *vista*. En el terreno de la medicina existen varios tipos de disfunción (visual, sexual, neuronal, etc.) que tienden a especificar más el concepto genérico proyectado por el hiperónimo y que a su vez pueden agrupar términos más específicos conformando una jerarquía donde el *genus* puede fungir como el nivel básico y las clases subsumidas ser categorías subordinadas que reflejen perspectivas de división relevantes de los hiperónimos en términos de que sean a) internamente cohesivas, b) externamente distintivas y c) maximalmente informativas (Cruse 2002).

Generalmente el modificador del hiperónimo localizado en definiciones analíticas es conceptualmente complejo debido a que vincula al hiperónimo con un conjunto de rasgos. Por ejemplo, en la categoría más específica *disfunción visual,* el adjetivo *visual* tiene asociados todo el conjunto de rasgos y propiedades del concepto representado por el nombre *vista*, que además se asocian composicionalmente con el hiperónimo *disfunción*. Esto contrasta con adjetivos calificativos (Demonte, 1999) o descriptivos (Fellbaum, 1998), tales como *grave* o *rara* que son conceptualmente simples y que difícilmente podrían considerarse como categorías más específicas con una carga de información

valiosa para el dominio (Croft y Cruse, 2004). Como se mencionó en la sección 2.4.4, los especialistas en un dominio específico tenderán a usar categorías subordinadas al *genus* o hiperónimo debido a que proveen de mayor información respecto al término definido. Así, resulta de gran interés la extracción de estas categorías utilizando como punto de partida los hiperónimos extraídos de los CDs analíticos y de otras fuentes, como son los primitivos semánticos indicadores de taxonimia, como lo es la expresión lingüística "tipo de" (Wierzbicka, 1996), que se presenta en la siguiente sección.

En resumen, en este trabajo nos enfocamos en extraer las categorías subordinadas del hiperónimo que sean más relevantes descartando los casos donde un hiperónimo se vincula con adjetivos que se relacionan con el contexto específico y no dan cuenta de categorías importantes para el dominio de acuerdo con la propuesta de Acosta *et al.*, (2010). Estos adjetivos pueden ser calificativos o descriptivos. Priorizamos los adjetivos relacionales como proveedores de un mayor número de propiedades por su origen nominal dada la observación de Cruse (2002) de que en muchas de las instancias donde un buen hipónimo no es un buen taxónimo de un hiperónimo, existe una descripción directa del hipónimo en términos del hiperónimo más un rasgo único. Un ejemplo ilustrativo que proporciona Cruse (2002) de esta situación es: *semental= caballo macho*, donde la división por género del hiperónimo *caballo* no provee de información significativa para la gran mayoría de las tareas de clasificación.

4.5.1 Extracción de hiperónimos a partir del primitivo semántico tipo de.

Primitivo semántico "tipo de"

Para Wierzbicka (1996), "tipo" (en inglés, kind) es un universal léxico que se encuentra en el núcleo de la categorización humana de los contenidos del mundo. De acuerdo con esta investigadora, el lexicón de todo lenguaje está repleto de categorías taxonómicas basadas fundamentalmente en este concepto. En inglés, por ejemplo, son relativamente comunes expresiones como: *A rose is a kind of flower* (en español, *una rosa es un tipo de flor*), *y oak is a kind of tree* (en español, *roble es un tipo de árbol*).

Wierzbicka señala que en el pasado se negó la importancia de la clasificación taxonómica basada en el concepto "tipo" en todos los lenguajes y culturas. Por ejemplo, se afirmó que en las sociedades no occidentales tradicionales, la clasificación etnobiológica básicamente era no taxonómica, es decir, no había ninguna idea de géneros, especies, familias o variedades; sin embargo, estas sociedades eran capaces de delinear los distintos objetos o elementos en su lenguaje (Lévy-Bruhl, 1926). En contraste, la clasificación científica occidental estaba basada en jerarquías de tipos.

Investigaciones posteriores llevadas a cabo por antropólogos y lingüistas revelaron que términos genéricos como árbol, pájaro, o pez, podrían ser poco usados en un lenguaje y, en su lugar, haber un mayor uso y abundancia de términos más específicos para denotar criaturas y plantas (Brown, 1984; Berlin, 1992). Sin embargo, la idea de que no existe una jerarquía de tipos en clasificaciones biológicas populares no ha perdurado a través del tiempo. De acuerdo con Wierzbicka (1996), lo anterior se sostiene básicamente por los siguientes puntos principales:

- La presencia universal de por lo menos alguna categorización jerárquica en el lexicón soporta la postura de que las taxonomías juegan un papel importante en la conceptualización de cosas vivientes. Por ejemplo, árbol-roble, pájaro-canario.
- En todo lenguaje conocido existe un conjunto de palabras que se consideran como los nombres reales de ciertas clases de cosas vivientes. Por ejemplo, cuando se le pregunta a alguien *¿cómo se llama esto?,* el informante podría contestar con un término genérico popular: *es como una lila,* o podría responder *no sé*, pero nunca dirá, *se llama ave*, o *se llama arbusto*. La presencia de estos nombres reales establece, más allá de cualquier duda razonable, la realidad psicológica de la noción de especies biológicas (o géneros populares).
- La evidencia lingüística sugiere que el concepto de "tipo" (o "tipos") es un universal léxico. Las oraciones que refieren a "tipos" de cosas vivientes son comunes (Goddard y Wierzbicka, 1994).

Finalmente, Wierzbicka (1996) concluye que el concepto "tipo" es un universal léxico debido a que no se puede reducir a algún otro concepto o

conceptos.

Análisis del concepto "tipo de" en el conjunto de definiciones de Wikipedia

De acuerdo con Cruse y Croft (2004), la expresión "tipo de" ejerce algún tipo de presión selectiva sobre el par de elementos vinculados, por lo que da margen a la existencia de un principio de subdivisión taxonómica que permite que se pueda predecir la relación de taxonimia de los significados de los elementos relacionados. Por su parte Wilks *et al.* (1996), asociados con definiciones analíticas encontradas en diccionarios, describen un conjunto de núcleos nominales que son indicadores de hiponimia, entre ellos el elemento *tipo* y añaden que el hiperónimo se encuentra generalmente después de la preposición "de".

Como se mencionó en párrafos anteriores, la expresión "tipo de" está vinculada estrechamente con definiciones analíticas, sin embargo, también se puede encontrar de forma independiente de una definición en cualquier discurso. Por ejemplo, en el caso de la siguiente oración:

> No todos los miembros de la familia van a tener exactamente el mismo tipo de estrabismo.

Aunque el elemento relacionado con *un tipo de estrabismo* se encuentra ausente, de la oración anterior podemos deducir que existen varios tipos de *estrabismo*, lo cual es el caso porque *exotropia, hipertropia e hipotropia* son tipos de *estrabismo*.

Para el caso del conjunto de CDs obtenido de Wikipedia, tenemos que la expresión "tipo" se encuentra relacionada con los siguientes nombres comunes:

Tipo de (receptor, glaucoma, crecimiento, virus, enunciado, adposición, gramática, etc.)

Dado lo anterior, presentamos una exploración de este primitivo semántico y núcleos relacionados (clase, forma) del conjunto de CDs de Wikipedia. El objetivo de esta heurística es generar más hiperónimos relevantes del corpus analizado. La tabla 18 muestra los resultados de esta exploración, estos resultados incluyen las ocurrencias en la FN del hiperónimo, así como también ocurrencias en la información de la diferencia específica.

Tabla 18. Hiperónimos relevantes.

Expresión	Hiperónimos relevantes	
	Frecuencia Absoluta	**Frecuencia porcentual**
Tipo <ADJ>? de <NC>	49	100%
Clase <ADJ>? de <NC>	4	80%
Forma <ADJ>? de <NC>	10	27%

Como lo muestran los resultados de la tabla 18, el núcleo que más ocurrencias presentó en los CDs es *tipo*. Además, es el más confiable para proporcionar buenos candidatos a hiperónimos. Así, se propone la exploración de este núcleo como fuente de hiperónimos relacionados con el dominio. Sin embargo, consideramos necesario proveer de mayor soporte empírico sobre su confiabilidad.

4.5.2 Recursos adicionales para la extracción de hipónimos

Las FNs que se utilizarán para extraer los hipónimos derivados de un hiperónimo se obtienen de una fase chunking que extrae las estructuras mostradas en la tabla 19. Además, se crean archivos que contienen la agrupación en frecuencias de las frases para efectos de cálculos posteriores con la medida de información mutua.

Tabla 19. Expresiones regulares de FNs.

Expresión regular	Ejemplo
<NC><ADJ>	Enfermedad cardiovascular
<NC><NC>[19]	Enfermedad Alzheimer

4.5.3 Medidas para determinar la asociación entre hiperónimos y modificadores

La metodología propuesta en este trabajo se enfoca en extraer un subconjunto de relaciones de hiponimia-hiperonimia a partir de CDs. Otro de nuestros intereses radica en estudiar las categorías más específicas (hipónimos) que se forman a partir del hiperónimo, por lo que una vez obtenidos estos hiperónimos, aplicamos una fase de *bootstrapping* en el mismo corpus con la

[19] Se incluyó esta estructura para determinar el porcentaje de NCs que realmente son adjetivos etiquetados erróneamente.

finalidad de extraerlos.

A partir del objetivo que nos hemos planteado es claro que requerimos de una medida o heurísticas que nos permitan obtener modificadores relevantes que den lugar a buenos candidatos a hipónimos. En este trabajo seleccionamos la medida de Información Mutua Puntual (IMP) dado su uso en tareas de extracción de colocaciones. En nuestro caso concreto, se exploró una variante normalizada de la medida IMP (Bouma, 2008), cuya normalización obedece a dos cuestiones fundamentales: usar medidas de asociación cuyos valores tengan una interpretación fija y reducir la sensibilidad a frecuencias bajas de ocurrencia de datos. La fórmula de IMP normalizada (IMPN) es la siguiente:

$$i_n(x,y) = \left(\ln \frac{p(x,y)}{p(x)p(y)} \right) \Big/ -\ln p(x,y)$$

En un intento por proporcionar una interpretación para los resultados IMP, Bouma señala que cuando dos palabras ocurren únicamente juntas, $i_n(x,y)=1$; cuando son independientes, $i_n(x,y) = 0$. Finalmente, cuando no ocurren juntas, $i_n(x,y) = -1$ y se aproxima a este valor cuando p(x,y) se acerca a 0 y tanto p(x) como p(y) permanecen fijos. Respecto al problema de la sensibilidad a frecuencias bajas, en la tabla 20 presentamos un conjunto de datos obtenidos del corpus Cli (Garduño *et al.*, 2004) donde se muestra la diferencia en valor que proporciona la medida IMPN comparada con la IMP tradicional, que proporciona un mismo valor para ambos casos. La relación que se evalúa es la de x = *empresa* con los modificadores adjetivos *maquiladora* y *filial*, ambas relaciones extraídas del corpus Cli.

Tabla 20. Comparación de medidas IMP.

Modificador adjetivo	**Frecuencia conjunta f(*empresa*, y)**	**Frecuencia marginal de *empresa***	**Frecuencia marginal de *y***	**IMP**	**IMPN**
Y = *maquiladora*	1	45	1	5.5	0.59
Y = *filial*	3	45	3	5.5	0.67

Algunas consideraciones importantes sobre los hiperónimos

Una de las observaciones importantes que es pertinente hacer en este punto es que los hiperónimos, como clases genéricas de un dominio, se espera que proyecten un número amplio de relaciones con modificadores adjetivos, nominales y de FP.

Como un ejemplo ilustrativo y sólo considerando el caso de los modificadores adjetivos, la figura 17 muestra datos del hiperónimo *enfermedad* y un subconjunto de 50 adjetivos más relacionados tomando en cuenta su información mutua puntual.

C(enfermedad, w_i)
Transmisible, prevenible, diarreica, diverticular, indicadora, autoinmunitaria, aterosclerótica, meningocócica, cardiovascular, pulmonar, afecto, febril, agravante, hepática, seudogripal, periodontal, sujeto, bacteriano, emergente, benigno, parasitaria, postrombotica, bacteriémica, coexistente, catastrófica, exclusiva, vectorial, supurativa, infecciosa, debilitante, digestiva, invasora, rara, inflamatoria, esporádica, antimembrana, predisponente, ulcerosa, contagiosa, cardiaca, sistémica, activa, grave, prexistente, miocárdica, somática, fulminante, atribuible, linfoproliferativa.

Figura 17. Adjetivos relacionados con el hiperónimo enfermedad.

Si seleccionamos un adjetivo relacional de la figura 17, por ejemplo, *cardiovascular* y otro calificativo, en este caso el adjetivo *raro*, y extraemos los nombres a los que modifican, obtenemos los datos de la figura 18. Ergo, tenemos que tanto el hiperónimo como el adjetivo, sea relacional o calificativo, pueden estar vinculados con otros elementos, situación que resta precisión a la medidas de asociación para detectar relaciones útiles.

$C(w_i$,cardiovascular)	$C(w_i$,rara)
efecto, problema, congreso, función, evento, relación, examen, inestabilidad, trastorno, enfermedad, bypass, causa, beneficio, sistema, reparador, descompensación, cirugía, operación, mortalidad, aparato, educación, síntoma, eficiencia, episodio, riesgo, investigación, manifestación, afección, medicamento, director, muerte, salud	televisión, enfermedad, complicación, infancia, niño, color, obesidad, mhc, nucleótido, sustancia, mutación, trastorno, grupo, meconio, epistaxis, derecha, síndrome, cáncer, alelo, forma, caso, párpado

Figura 18. Nombres modificados por los adjetivos cardiovascular y raro.

De acuerdo con Demonte (1999), los adjetivos relacionales y los nombres que modifican se relacionan composicionalmente, es decir, en el caso de una construcción como *enfermedad cardiovascular*, el adjetivo relacional *cardio-vascular* vincula todos los rasgos o propiedades del *corazón* y *vasos sanguíneos* al nombre *enfermedad*. Dada esta situación, proponemos explorar heurísticas que se relacionan con pruebas lingüísticas que propone Demonte (1999) para discernir entre adjetivos calificativos y relacionales. Dichas heurísticas se describen en la siguiente sección.

4.5.4 Heurísticas lingüísticas para la recuperación de hipónimos relevantes

Una de las propuestas de nuestro trabajo es la extracción de hipónimos que tengan como núcleo un hiperónimo. La estructura de estos hipónimos es aquella de los patrones más comunes de construcción de términos en español, es decir, un núcleo nominal con modificadores adjetivos (Vivaldi, 2001). Por ejemplo, a partir de un hiperónimo como enfermedad, podemos obtener: *enfermedad mitocóndrica*, *enfermedad cardiovascular*.

Para lograr lo anterior, el conjunto de hiperónimos extraído de los CDs se utiliza como *semilla* para obtener sus elementos asociados de las FNs mencionadas en la sección 4.5.2. Esta sección describe una serie de heurísticas lingüísticas propuestas por Demonte (1999) que tienen como objetivo discernir entre adjetivos calificativos y relacionales. El énfasis de este

trabajo es proponer un filtrado de modificadores adjetivos que nos permitan obtener una alta precisión con una cobertura aceptable, dada la composicionalidad que guardan los adjetivos relacionales con los hiperónimos que modifican.

El adjetivo: un elemento importante para la construcción de términos

De acuerdo con Demonte (1999), "el adjetivo es una categoría gramatical que modifica al nombre: una clase de palabras cuyos miembros tienen características formales muy precisas y es también una categoría semántica: hay un tipo de significado que se expresa preferentemente por medio de adjetivos."

Los adjetivos representan un elemento muy importante en la construcción de términos, por tal motivo en esta investigación resulta de gran interés realizar una revisión general de sus características en aras de proveer de una tipificación lingüística adecuada que se apoye con datos obtenidos de corpus y que nos permita filtrar aquellos adjetivos que son valorativos o descriptivos en situaciones específicas y que con frecuencia no participan en la construcción de términos, por ejemplo: *enfermedad grave*.

Retomando el trabajo de Demonte (1999), existen dos clases de adjetivos que asignan propiedades a los nombres. La diferencia entre estos dos tipos consiste en el número de propiedades que cada uno conlleva y la manera como la vinculan con el nombre. Por un lado, están los adjetivos que refieren a un rasgo constitutivo del nombre modificado, rasgo que exhiben o caracterizan a través de una única propiedad física: color, forma, carácter, predisposición, sonoridad (ver tabla 21):

- Libro azul
- Señora delgada
- Hombre simpático

Tabla 21. Clases de adjetivos calificativos.

Tipo de adjetivo	Ejemplos
Dimensión	largo, alto, corto, bajo, ancho, amplio, angosto, grueso, fino, grande, etc.
Propiedad física de objetos perceptibles mediante los sentidos	redondo, ovalado, etc.
color y forma	blanco, negro, gris, rojo, verde, rojizo, azulado, verde botella, etc.
De tiempo o edad	viejo, nuevo, joven, antiguo, arcaico, lejano, reciente, moderno, añejo, etc.
Valoración o evaluativos	bueno-malo, lindo-feo, bello, bonito, perfecto, horrible, tremendo, pésimo, etc.
Aptitudes y (pre)disposiciones humanas intelectuales	inteligente, capaz, sabio, despierto, astuto, sagaz, idiota, memo; emocionales: sensible, amable, cordial, simpático, etc.

Por otro lado, se encuentran los adjetivos que se refieren a un conjunto de propiedades (todas las características que, conjuntamente, definan a un nombre como *mar*, *leche* o *campo*) y las vinculan a las del nombre modificado:

a) Puerto marítimo
b) Vaca lechera
c) Paseo campestre

Dado lo anterior, nuestra propuesta consiste en enfocar la atención en los adjetivos relacionales y, para discernir éstos de aquellos calificativos, aplicamos tres criterios propuestos por Demonte (1999), con el objetivo en mente de que los datos que puedan ser de utilidad para discernir entre estas dos clases se obtengan de la misma fuente de entrada, sin estar supeditados a una lista de paro fija, dado el estatus de clase abierta que tiene la categoría gramatical de los adjetivos. Las heurísticas, a grandes rasgos, son las siguientes:

- La posibilidad de que un adjetivo sea usado predicativamente. Ejemplo, *El método es importante*.
- El que un adjetivo sea parte de comparaciones, de modo que su significado sea modificado por adverbios de grado: *relativamente rápido*.

- La precedencia del adjetivo respecto al nombre: *Una grave enfermedad*.

Finalmente, la vinculación de adjetivos relacionales con nombres con los que comparten contenido semántico es muy relevante en aplicaciones tales como WordNet (Fellbaum, 1998). En el caso de WordNet, los adjetivos se vinculan mediante una red de antónimos, sin embargo, al no tener antónimos obvios la mayoría de los adjetivos relacionales, simplemente se ligan con el nombre con el que comparten contenido semántico.

4.6 Resumen de la metodología propuesta

La figura 19 muestra el esquema general de la metodología descrita a lo largo de este capítulo. A continuación presentamos un resumen del proceso:

1. Preprocesamiento de la fuente textual de entrada. Consiste en la eliminación de algunos elementos de la colección de textos (tablas, direcciones de correo electrónico, información en paréntesis, etc.). Aunado a lo anterior, se realiza un proceso de segmentación de oraciones previo al etiquetado de partes de la oración. Después de la fase de etiquetado de partes de la oración, se aplica un proceso de corrección de lemas y etiquetas. Finalmente, se realiza la normalización de un subconjunto de etiquetas.

2. Fase de extracción conceptual. Consiste en la aplicación de una gramática de expresiones regulares que considera el comportamiento sintáctico más común de los elementos de un CD, todo esto en el marco de un patrón contextual específico. Posteriormente, los candidatos a CDs, resultado del filtro sintáctico, constituyen la entrada a un proceso adicional de filtrado de nombres indicadores de relaciones de meronimia-holonimia y causalidad.

3. Fase de extracción de términos e hiperónimos. La etapa de extracción de hiperónimos se basa en el filtrado de núcleos vacíos o confusos, es decir, cuando se localiza un núcleo vacío

se procede a la búsqueda del hiperónimo más relacionado semánticamente tomando en cuenta la etiqueta gramatical. Además, en los casos donde el hiperónimo está ausente pero existe una expresión focalizando el núcleo nominal de la frase correspondiente al término, se considera que este núcleo es el hiperónimo. Finalmente, el resto de los casos simplemente se filtra a partir de su etiqueta gramatical, es decir, se extrae el primer nombre que ocurra en la FN del hiperónimo.

4. Fase de extracción de hipónimos a partir de hiperónimos. Los hipéronimos asociados con modificadores conceptualmente complejos se consideran como categorías subordinadas o hipónimos, que son valiosos para el dominio, por lo que esta última etapa se enfoca en su extracción. Para lograr lo anterior, el conjunto de hiperónimos más frecuentes extraído de los CDs, o del primitivo semántico “tipo de”, o de una combinación de ambos, se utiliza para extraer estos modificadores. La fuente para extraer estos modificadores es el conjunto de FNs obtenida de una fase chunking del mismo corpus de entrada.

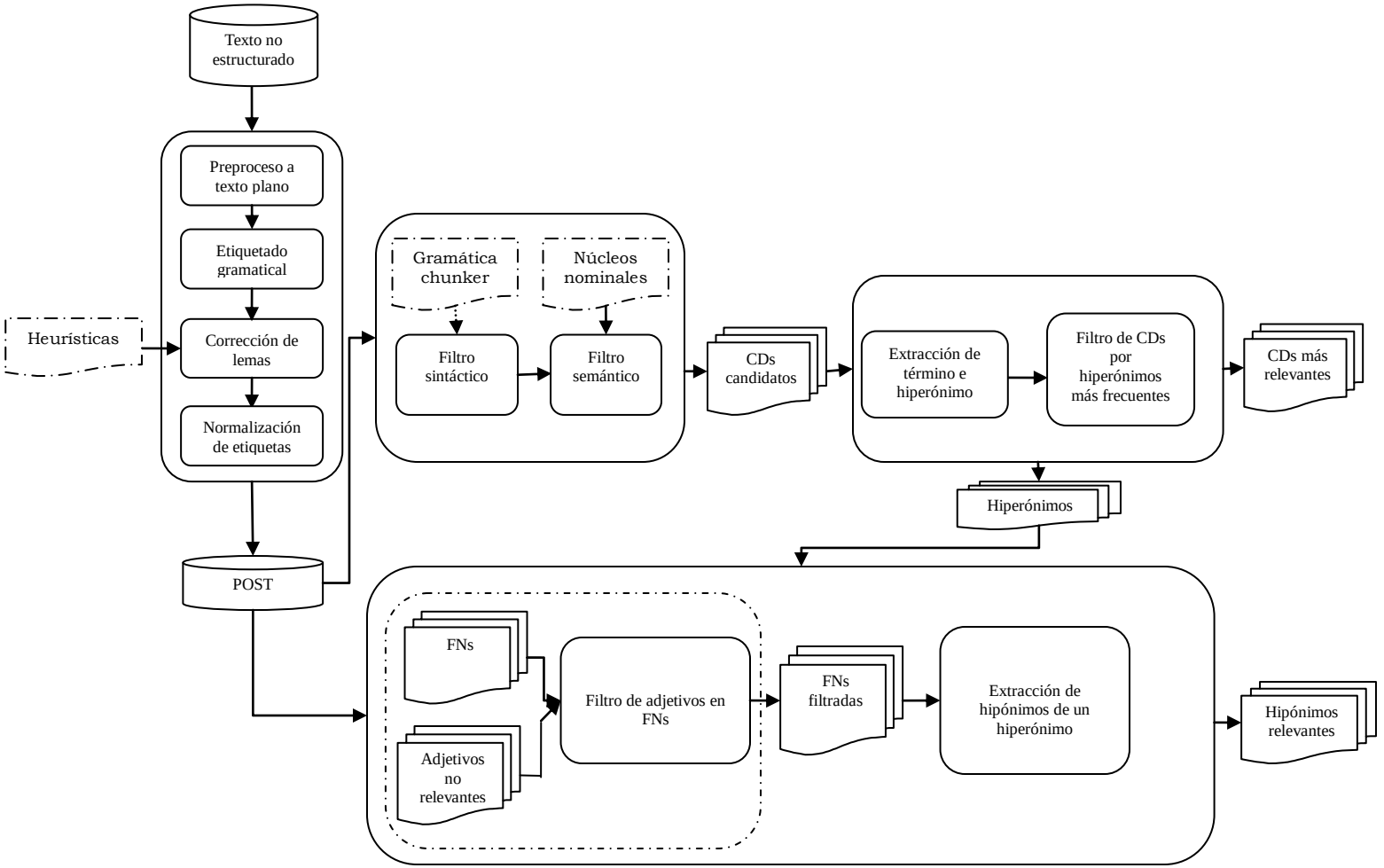

Figura 19. Metodología para la extracción de relaciones de hiponimia-hiperonimia.

Capítulo 5

Implementación y análisis de resultados

El objetivo de este capítulo es presentar los resultados de la aplicación de la metodología propuesta en nuestra investigación a una colección de textos específica. En primera instancia, describimos a grandes rasgos las colecciones de textos consideradas, así como también las herramientas que se utilizaron para automatizar cada uno de los procesos que contempla la metodología. En segundo término, se describen los resultados de la fase de extracción conceptual donde se evalúa el desempeño de nuestra metodología con un *baseline* establecido *a priori* y con la herramienta Ecode sólo para el caso de CDs analíticos. Posteriormente, se muestran los resultados de la extracción de hiperónimos de los fragmentos de CDs, en términos de la precisión y la cobertura respecto a los hiperónimos presentes en la colección de textos, así como también los resultados del filtro de hiperónimos más frecuentes. Otra sección adicional da cuenta de los resultados en la etapa de extracción de hipónimos a partir de hiperónimos donde se compara la medida de información mutua estandarizada con heurísticas lingüísticas para extraer hipónimos relevantes. Finalmente, se implementa una herramienta de extracción de relaciones de hiponimia vía el aprendizaje de patrones e instancias considerando un subconjunto de las relaciones extraídas de CDs.

5.1 Recursos

5.1.1 Selección de la fuente de información textual

La fuente de información textual está constituida por un conjunto de documentos del dominio médico, básicamente enfermedades del cuerpo humano y temas relacionados (cirugías, tratamientos, estudios, etc.). Estos documentos se obtuvieron de MedLine Plus en español. Los documentos colectados corresponden a dos géneros diferentes, por un lado información más cercana a conocimiento enciclopédico respecto al tema de interés donde regularmente se puede encontrar una definición y más detalles respecto al concepto definido. Por otro lado, noticias de salud relacionadas con alguna enfermedad, tratamiento o estudio. Finalmente,

se recolectaron cuatro libros en formato PDF del mismo dominio médico, con temáticas tales como medicina interna, enfermedades transmisibles, podología geriátrica, y prevención y control de enfermedades.

La información disponible en el sitio de MedLine Plus va dirigida de expertos a (semi)expertos (pacientes, principiantes del área médica, etc.) y se actualiza continuamente. Del mismo modo, la información de los libros de texto está enfocada por lo menos a principiantes en el área de la medicina.

El tamaño total del conjunto de documentos de medicina es de 1.3 millones de palabras. Seleccionamos un dominio médico por razones de disponibilidad de recursos textuales en formato digital. Sin embargo, asumimos que la selección de este dominio no supone restricciones muy fuertes para la generalización de resultados a otros dominios.

5.1.2 Relaciones de hiponimia-hiperonimia presentes en la fuente de información textual

Como se mencionó en la sección 1.6.4, la extracción conceptual solo considera los verbos de la tabla 5, que de acuerdo con Alarcón (2009) son característicos de definiciones analíticas. En una revisión semi-automática del corpus de medicina se encontraron las relaciones que se presentan en la tabla 22. Se consideró importante clasificar las relaciones por su origen, es decir, de acuerdo al tipo de expresión donde se encuentran implícitas. Por ejemplo, existen 1273 CDs analíticos en el corpus, los que a su vez tienen implícitas la misma cantidad de relaciones de hiponimia-hiperonimia. Por otro lado, otros patrones tales como: *Xs incluyendo Ys y Zs*, tienen implícitas 1710 relaciones. Existe un subconjunto de 92 relaciones en común entre los CDs y otros patrones, por lo que mostramos un subtotal donde estos elementos se han eliminado. Finalmente, los hipónimos derivados de hiperónimos más modificadores adjetivos o nombres generan un conjunto de 4758 relaciones. Finalmente, la estructura hiperónimo + FP (preposición “de”) genera un total de 4476 relaciones[20].

[20] Es importante mencionar en este punto que no se considera la variación de permutación mencionada en Vivaldi (2001) al contabilizar el número de relaciones. Es decir, enfermedad renal y enfermedad del riñón, se consideran como dos relaciones diferentes.

Tabla 22. Relaciones de hiponimia-hiperonimia presentes en corpus.

Relaciones de hiponimia-hiperonimia	Corpus de medicina
CDs	1273
Otros patrones	1710
Subtotal:[21]	2891
Hiperónimo+adjetivo	4758
Hiperónimo+FP	4476
Total:	12125

5.1.3 Herramientas computacionales

El lenguaje de programación usado para automatizar todas las tareas requeridas fue Python versión 2.6, así como también el módulo de procesamiento de lenguaje natural NLTK. El módulo NLTK constituye un recurso muy valioso para la investigación y el desarrollo de herramientas de procesamiento de lenguaje natural.

5.2 Resultados de la fase de extracción conceptual

La evaluación de la etapa de extracción conceptual se realiza considerando como *baseline* los candidatos a CDs obtenidos de sólo filtrar por medio de los verbos más característicos de definiciones analíticas e imponer una restricción adicional en términos de la longitud de oración de 4 palabras o mayor.

Aunado a lo anterior, se evalúan los niveles de precisión y cobertura obtenidos mediante la herramienta Ecode (sólo para el caso de CDs analíticos) y la metodología propuesta en este trabajo. La precisión se define como el cociente de candidatos relevantes obtenidos en el proceso de extracción dividido por el número total de casos recuperados en el proceso de extracción:

$$P = \left(\frac{CasosRelevantesExtraídos}{\#TotaldeCasosExtraídos} \right)$$

Por otro lado, la cobertura corresponde al cociente derivado del número de casos relevantes recuperados en el proceso de extracción dividido por el total de casos relevantes en el corpus:

$$C = \left(\frac{CasosRelevantesExtraídos}{\#TotaldeCasosRelevantesEnCorpus} \right)$$

[21] El subtotal excluye las relaciones que se encuentran en los CDs y en otros patrones.

Finalmente, la medida F que se considera en este trabajo es la que da el mismo peso a la precisión (P) y a la cobertura (C):

$$F = \left(2 * \left(\frac{P * C}{P + C}\right)\right)$$

Los resultados derivados de nuestra metodología se presentan en la tabla 23. Se realizó la comparación con los resultados de la herramienta Ecode, sólo para el caso de definiciones analíticas. Como se mencionó en la sección 1.6.4, Ecode extrae CDs de fuentes textuales y estos CDs no solo se restringen a analíticos, sino también a sinonímicos, funcionales y extensionales (Sierra *et al.*, 2008).

Tabla 23. Resultados de la extracción conceptual.

	Medicina		
	P	**C**	**F**
Baseline	8%	100%	15%
Ecode	24%	63%	35%
Filtro sintáctico	60%	56%	58%
Filtro semántico	68%	56%	61%

De la tabla 23 se puede observar que nuestra metodología tiene un desempeño mejor en cuanto a precisión se refiere que el extractor conceptual Ecode. Sólo considerando el filtro sintáctico se logra una precisión (P) de 60% y una cobertura (C) del 56%, que está por debajo del 63% obtenido por Ecode, sin embargo, priorizamos la precisión debido a que el objetivo es la obtención automática de un subconjunto *semilla* de relaciones de hiponimia-hiperonimia confiable de los CDs.

Los resultados con el filtro semántico muestran una mejora en la precisión, que aumenta un 8% sin afectar de forma significativa la cobertura, que se mantiene en un 56%. En el apéndice se incluyen las tablas con los hiperónimos candidatos obtenidos del corpus de medicina sólo para frecuencias de ocurrencia mayores o iguales a 3.

5.2.1 Consideraciones importantes

Filtro sintáctico

El filtro sintáctico constituye una fase que resulta importante para garantizar la existencia de un término, con una estructura específica, y una FN donde sea posible localizar el hiperónimo correspondiente, ambos elementos vinculados con un patrón verbal particular. Sin embargo, la estructura sintáctica no es suficiente para filtrar la totalidad de CDs relevantes debido a que existen contextos que pueden llegar a ser muy complejos y las causas de esta complejidad ser variable. Aunado a lo anterior, muchos fragmentos, que no son CDs, tienen una estructura similar, sobre todo aquellos con el verbo *ser*, que es el más productivo.

A partir de los resultados obtenidos con el filtrado sintáctico podemos afirmar que la regularidad sintáctica de los CDs es por lo menos de un 50%, lo que podríamos considerar corresponde al comportamiento más canónico. El porcentaje restante corresponde a CDs que tienen un comportamiento más complejo. Los siguientes ejemplos se incluyen con la intención de ilustrar mejor parte de esta complejidad. Además, se proporciona una paráfrasis de esta información en términos de un CD analítico más cercano a lo canónico.

1. El conjunto de aparatos puestos al servicio del sistema eléctrico, que vigilan que se cumpla adecuadamente el propósito para el que fue creado, es lo que se conoce como protección. **Paráfrasis canónica**: Protección es el conjunto de aparatos puestos al servicio del sistema eléctrico, que vigilan…
2. Las pruebas que generalmente se efectúan a los interruptores o antes de poner en servicio un sistema, son las siguientes: prueba de prestación. **Paráfrasis canónica**: Pruebas de prestación son las pruebas que generalmente se efectúan a los interruptores o antes de poner en servicio un sistema.
3. En bebés o en niños pequeños, se puede utilizar un instrumento puntiagudo llamado lanceta para punzar la piel y hacerla sangrar. **Paráfrasis canónica**: Una lanceta es un instrumento que se utiliza para punzar la piel y hacerla sangrar.
4. La sangre se recoge en un tubo pequeño de vidrio llamado pipeta, en un portaobjetos o en una tira reactiva. **Paráfrasis canónica**: Una pipeta es un

tubo pequeño de vidrio que sirve para recoger sangre.

5. El mejor método para efectuar los cálculos con corrientes desequilibradas de falla en grandes sistemas de energía es el conocido como componentes simétricas. **Paráfrasis canónica**: Componentes simétricas es el mejor método para efectuar los cálculos con corrientes desequilibradas…

Como podemos observar de los CDs anteriores, la posición de los elementos de un CD pueden encontrarse invertidos, como es el caso de 1, 2 y 5, donde *aparato, prueba* y *método* son los hiperónimos de *protección, pruebas de prestación* y *componentes simétricas* (sus hipónimos), respectivamente. Por su parte, la *differentia* se localiza entre el hiperónimo y el hipónimo, lo que no se captura con los patrones considerados en la gramática. Tomar en cuenta estos patrones implicaría la obtención de una mayor cantidad de ruido que superaría significativamente la ganancia en cobertura, por lo que consideramos, por el momento, no es adecuada su inclusión.

Para el caso de 3 y 4, el hecho de encontrar el fragmento más relevante (donde se encuentra el término y la FN del hiperónimo) inserto en una oración mayor, con una gran variedad de elementos vecinos del hiperónimo, obligó a la restricción por medio de un punto (“.”) o coma (“,”) antes de la posición del hiperónimo (ver tabla 14 de la sección 4.3.1), todo esto con la finalidad de disminuir el ruido en los resultados.

Por otro lado, como se mencionó en la sección 4.2.5, los etiquetadores de partes de la oración disminuyen su eficacia en dominios especializados debido a que generalmente se entrenan con corpus de lengua general, o bien, con corpus de un dominio específico para luego aplicarse en otro. Concretamente para el etiquetador TreeTagger, sin algún tipo de entrenamiento, el 24% de oraciones presentaron errores de etiquetado. Los errores más comunes son precisamente aquellos que involucran nombres y adjetivos, particularmente: adjetivos que se etiquetan como nombres y nombres que se etiquetan como verbos, que conjuntamente representan un porcentaje estimado del 52% (ver tabla 9, capítulo 4). Estos tipos de error sin duda impactan negativamente en los resultados debido a que impiden la extracción de los elementos relevantes con los patrones considerados.

Filtro semántico

Como se mencionó en la sección 4.3.2 de la metodología, los verbos utilizados para construir definiciones analíticas no sólo se utilizan para este fin. Uno de los casos más ilustrativos es el verbo "ser". Para disminuir el ruido generado por la inclusión del verbo *ser* se consideró un filtro semántico que removiera las relaciones parte-todo y causalidad detectadas a partir de núcleos nominales en la FN del término o del hiperónimo, que son relativamente comunes en esquemas analíticos. Este filtro semántico adicional logra aumentar los niveles de precisión sin afectar significativamente la cobertura, es por ello que resultó ser de gran utilidad.

5.3 Resultados de la fase de extracción de hiperónimos

A grandes rasgos, la fase de extracción de hiperónimos de los CDs se realizó considerando las heurísticas discutidas en Wilks *et al.* (1996) respecto a núcleos nominales vacíos o confusos, la identificación de pronombres relativos en lugar del *genus*, así como también del filtro de etiquetas gramaticales. Los resultados se muestran en la tabla 24. Además, como se mencionó en la sección 4.4, si en la etapa de extracción de hiperónimos no se extrae algún elemento, entonces el CD se elimina del conjunto, por ello etiquetamos estos indicadores de precisión y cobertura como "desempeño después de la extracción de hiperónimos". En el caso particular del corpus analizado, esta etapa filtró un total de 27 CDs no relevantes, lo que aumenta la precisión a un 70% y la cobertura permanece en un 56%.

Tabla 24. Fase de extracción de hiperónimos de CDs.

	Desempeño después de la extracción de hiperónimos		**Extracción de hiperónimos**		**Extracción de hipónimo e hiperónimo**	
	P	**C**	**P**	**C**	**P**	**C**
Medicina	70%	56%	96%	34%	96%	6%

Los datos mostrados en la sección 5.2 reflejan la precisión y la cobertura global de la extracción conceptual. Como se recordará, la precisión de esta fase con el filtro semántico es de un 68%. De este porcentaje de CDs relevantes, la

extracción de hiperónimos e hipónimos tiene una precisión del 96%, siendo aquellos donde se encuentra invertido el término y el hiperónimo los que impiden una precisión del 100%. Un ejemplo de lo anterior es el siguiente:

El antibiótico más indicado como agente profiláctico es la rifampicina.

En el ejemplo anterior la construcción *más indicado como agente profiláctico* es cubierto por una estructura de sinónimos. Como se recordará, consideramos dos estructuras sintácticas para este constituyente: Sin1 y Sin2. El fragmento de CD se extrae vía un patrón contextual con el verbo *ser*, por lo que el sistema extrae el término de la frase anterior al verbo *ser* y el hiperónimo de la frase posterior al verbo, lo cual es erróneo. La cobertura únicamente de hiperónimos es del 34% y la correspondiente al par hipónimo-hiperónimo es del 6%. Esta última cobertura es baja, sin embargo, nuestro objetivo no es extraer la mayor cantidad de relaciones posibles, sino un subconjunto confiable que sirva como *semilla* en procesos de aprendizaje automático de patrones e instancias de la relación.

Hiperónimos más frecuentes como filtro de candidatos a CDs

Como se mencionó en la sección 4.4, en esta investigación proponemos el uso de los hiperónimos más frecuentes para filtrar CDs analíticos no relevantes. Los resultados que se obtuvieron de este proceso se presentan en la tabla 25. Sólo se establecen 3 umbrales de frecuencias y se analizan los indicadores de precisión, cobertura y medida F. De la tabla es claro que a medida que el umbral de frecuencia es mayor se tiene una mejor precisión. Por ejemplo, para el caso de un umbral de ocurrencia del hiperónimo mayor que 20, se obtiene una precisión muy alta del 94%, sin embargo, la cobertura es baja (15%).

Tabla 25. Umbral de frecuencia de hiperónimos para el corpus de medicina.

Umbral	P	C	F
Frec>20	94%	15%	26%
Frec>10	86%	26%	40%
Frec>5	82%	37%	51%

5.3.1 Primitivo semántico “tipo de” y sus variantes

El primitivo semántico “tipo de” resultó ser el más confiable para obtener hiperónimos, no sólo en el contexto de definiciones, sino en el discurso en

general. Con el primitivo “tipo de” se obtuvo una precisión (P) de 84% y una cobertura (C) del 33%. En lo que respecta a la expresión “clase de”, se obtiene una confiabilidad alta, sin embargo su productividad es baja. Finalmente, la expresión “forma de” hace alusión principalmente a la forma o apariencia de cosas u objetos. La tabla 26 muestra los resultados obtenidos.

Tabla 26. Hiperónimos relevantes y no relevantes.

Construcción Lingüística	**Indicadores de desempeño**		
	P	**C**	**Hiperónimos relevantes**
Tipo de <NC>	84%	33%	236/283
Forma de <NC>	54%	16%	112/209
Clase de <NC>	85%	3%	23/27

Por otro lado, el porcentaje de traslape o de hiperónimos comunes[22] obtenidos vía los CDs y la expresión “tipo de” es de un 13%. Tomando en cuenta ambas fuentes de hiperónimos tenemos una cobertura de hiperónimos del 59% y una precisión del 77%. Lo anterior nos da la pauta para proponer que ambas fuentes, CDs y primitivo semántico ”tipo de”, pueden utilizarse como complementarias para la obtención de hiperónimos. La tabla 27 resume los resultados descritos anteriormente.

Tabla 27. Evaluación de la extracción de hiperónimos de CDs y primitivo “tipo de**”.**

Fuente	**Medida**	**Indicadores de desempeño**
CDs	Cobertura	34%
	Precisión	70%
	Medida F	46%
Primitivo “Tipo de”	Cobertura	33%
	Precisión	84%
	Medida F	47%
Conjunta	Cobertura	59%
	Precisión	77%
	Medida F	67%
Traslape		13%

[22] Calculado a partir de dividir el subconjunto de hiperónimos comunes entre la suma de hiperónimos extraídos vía los CDs y la expresión “tipo de”.

En resumen, los resultados obtenidos soportan la afirmación de Wierzbicka (1996) relacionada con el primitivo "tipo de" como característico de hiponimia, por lo que puede usarse como una fuente más para obtener buenos candidatos a hiperónimos, además de ser una fuente complementaria a los obtenidos de los CDs, donde se puede aplicar también el establecimiento de umbrales de ocurrencia para obtener los más confiables.

5.3.2 Tipos de hiperonimia encontrados en CDs

Una de las ventajas importantes de extraer relaciones de hiponimia-hiperonimia de CDs analíticos es la obtención de más de una categoría asignada al concepto definido. Como se mencionó en el capítulo 2, estas categorías pueden ser variantes léxicas, como el caso de *inflamación-hinchazón*. En otros casos, si asumimos la existencia de un sistema conceptual jerarquizado, estas categorías podrían estar en diferentes niveles de un sistema conceptual: dietista-profesional, dietista-persona. Finalmente, estas categorías podrían reflejar la falta de un consenso en cuanto a la categoría a la que pertenece un concepto; ejemplos dentro del dominio de la Biología abundan, sin embargo, Smith y Medin (1981) señalan como ilustrativo el concepto *Euglena.*

El capítulo 2 se enfocó en una presentación general de lo que son los conceptos desde una perspectiva psicológica. Estos contenidos son útiles para enmarcar parte del comportamiento que esperamos en definiciones analíticas al analizar grandes volúmenes de información textual. En algunos casos, la información extraída de los CDs, en cuanto a hiperónimos se refiere, reflejó precisamente este tipo de situaciones. Por ejemplo, en el corpus de medicina es muy claro el comportamiento de los términos con sufijo "itis" que llevan el significado de *hinchazón* e *inflamación* de algún órgano. Ambos se localizaron categorizando conceptos en definiciones analíticas, sin embargo, el más utilizado parece ser *inflamación* porque derivó en un mayor número de ocurrencias. Un par de ejemplos son los siguientes:

a) Uveítis es la hinchazón e irritación de la úvea, la capa media del ojo que suministra la mayor parte del flujo sanguíneo a la retina.
b) Neuritis óptica es la inflamación del nervio óptico que puede causar una reducción repentina de la visión en el ojo afectado.

Por otro lado, encontramos casos donde la categoría asignada al concepto se encuentra en algún nivel de una jerarquía conceptual. Los siguientes ejemplos ilustran esta situación:

c) Un dietista es una persona especialmente capacitada para planificar dietas saludables.
d) Un dietista es el profesional que se encarga de estudiar, vigilar y recomendar los hábitos alimenticios con el objetivo de mantener o mejorar su salud.

En el escenario anterior podemos observar que la categoría *persona* se encuentra en un nivel superordinado a la categoría *profesional*. Dado los resultados derivados de la psicología cognitiva, esperaríamos que dentro de dominios especializados fuese más común la utilización de categorías más específicas para clasificar términos o conceptos del dominio. Sin duda, resultados de la aplicación de nuestra metodología a grandes volúmenes de información textual pueden aportar más soporte empírico para este tipo de afirmaciones.

Respecto a la falta de consenso en la clasificación de un concepto, aunque no localizamos casos ilustrativos como el mencionado por Smith y Medin (1981) para el concepto *Euglena*, suponemos que son posibles dentro de dominios especializados, sobre todo cuando se analizan fuentes de gran tamaño y de diversos autores sobre un mismo dominio. Con el objetivo de ilustrar una situación que llamó nuestra atención respecto a diferentes categorías asociadas a un mismo concepto encontrados en diferentes contextos, buscamos información sobre el concepto *diabetes* en la Web y encontramos las siguientes definiciones:

e) La diabetes mellitus (DM) es un conjunto de trastornos metabólicos, que afecta a diferentes órganos y tejidos, dura toda la vida y se caracteriza por un aumento de los niveles de glucosa en la sangre: hiperglucemia.
f) La diabetes es una enfermedad crónica relacionada con un nivel alto de azúcar en la sangre que puede causar la pérdida de la visión.
g) La diabetes es un desorden del metabolismo, el proceso que convierte el alimento que ingerimos en energía.

Del escenario anterior podríamos cuestionar: ¿hasta qué punto trastorno-enfermedad-desorden son sinónimos?, o ¿representan realmente clases diferentes? Consideramos que este tipo de diferencias tal vez no reflejan una falta de consenso real como el mencionado por Smith y Medin (1981), sino simplemente categorías que se consideran sinónimas en un contexto específico.

Finalmente, y no por ello menos importante, es la presencia de nominalizaciones de un verbo en la posición del hiperónimo. Por ejemplo, para el término *arteriopatía coronaria* extraemos la definición h) que enfatiza lo que provoca o causa esta enfermedad; sin embargo, no menciona la categoría genérica en la que se clasifica el concepto, en este caso particular, como lo refleja la definición en i), se trata de una *enfermedad*. Además, otro tipo de información que complementa esta nominalización es el argumento *vasos sanguíneos*. Dado lo anterior, podríamos asumir que más que representar una definición analítica, el caso de la definición en h) lleva implícita una relación de causalidad, donde *arteriopatía coronaria* es el agente y *vasos sanguíneos* el tema de un predicado como *estrechar*.

h) La arteriopatía coronaria es un estrechamiento de los pequeños vasos sanguíneos que suministran sangre y oxígeno al corazón.
i) La arteriopatía coronaria es un tipo de enfermedad cardíaca que causa un suministro inadecuado de sangre al músculo cardíaco, una afección potencialmente perjudicial.

Casos como el anterior se *disfrazan* en contextos analíticos y, de acuerdo con nuestro análisis, pueden representar varios tipos de información. Para el caso concreto del corpus de medicina observamos que las nominalizaciones hacen referencia a relaciones de causalidad, donde el término definido puede ser el agente y el argumento de la nominalización (si está presente), el tema o paciente. Dicho argumento generalmente está localizado después de la preposición "de" (estrechamiento de los vasos sanguíneos). En este sentido, en Sierra *et al.*, (2012) se intenta dar cuenta del tipo de relaciones que proyectan los hiperónimos cuando una FP preposicional con núcleo "de" se encuentra presente. A continuación se muestra otro caso ilustrativo:

j) El esputo es una secreción que se produce en los pulmones y en los bronquios que puede ser expulsada cuando se da una tos profunda.
k) Esputo es una sustancia secretada por las vías respiratorias que se arroja por la boca de una vez.
l) Se llama secreción al proceso por el que una célula o un ser vivo vierte al exterior sustancias de cualquier clase.

La definición en j) está más acorde con una relación de causalidad donde el concepto *esputo* es el resultado del proceso de *secretar*, y el agente son los *pulmones* y los *bronquios*. Por su parte, la definición k) clasifica el concepto como una *sustancia* y, adicionalmente, complementa en la diferencia específica la relación de causalidad. La intención de incluir la definición l) es para dar cuenta que *secreción* se considera como un proceso y no como *materia* o *sustancia*, que es lo que denota el concepto *esputo.*

Antes de concluir esta sección es importante mencionar que Fellbaum (1998), en el marco del diseño y construcción del lexicón WordNet, señala que se dieron cuenta de que el apuntador de hiponimia representaba más de una relación de hiponimia después de haber diseñado la jerarquía de nombres (Fellbaum, 1998). Sin embargo, actualmente se encuentra contemplada esta distinción vía la configuración de *synsets* diferentes para cada sentido, cada uno con sus relaciones específicas. Para dar cuenta de lo anterior, Fellbaum se basó en el trabajo de Wierzbicka (1984), quien propuso cinco tipos de relaciones hiponímicas, dos de las más relevantes son: IS-A-KIND-OF e IS-USED-AS-A-KIND-OF. La primera de ellas es la más discutida en la literatura para representar relaciones taxonómicas. La segunda tiene que ver con aspectos funcionales y se considera generalmente para el caso de artefactos. Por su parte Pustejovsky (1991) propone la estructura Qualia para representar el significado y polisemia de elementos léxicos. En el marco de la estructura Qualia se contempla un rol formal y otro télico, que se corresponden con el taxonómico y funcional de Wierzbicka. Estos dos trabajos resultan importantes precisamente porque logran explicar estas diferencias en la relación de hiponimia que se pueden encontrar comúnmente en la formulación de contenido conceptual en términos analíticos. Es importante mencionar que la no consideración de este tipo de distinciones implica problemas en un escenario de representación de conocimiento, y

considerarlas contribuye a dar cuenta de la polisemia inherente en las unidades léxicas, así como también de su significado (Pustejovsky, 1991).

Por último, los ejemplos mostrados en párrafos anteriores evidencian un rol de tipo agentivo, donde el hiperónimo proyecta una relación que puede interpretarse como el hipónimo siendo la causa en el caso del CD: la arteriopatía coronaria es un estrechamiento de los pequeños vasos sanguíneos. O bien, la consecuencia, que es el caso del CD: esputo es una secreción. Lo anterior se puede explicar en parte porque en estos casos no se trata de artefactos, sino de procesos o eventos.

5.4 Resultados de la extracción de hipónimos a partir de hiperónimos

En definiciones analíticas, un concepto se define en términos de un *genus* y una o más *differentiae*. Con frecuencia el *genus* se vincula con modificadores, ya sean adjetivos, nombres comunes o FPs, que forman parte de la *differentiae* y que contribuyen a formar categorías subordinadas o más específicas que contienen información valiosa para los expertos de un dominio particular, como se mencionó en la sección 2.4.4. Dado lo anterior, el objetivo de este apartado es explorar el potencial de los *genus* o hiperónimos más frecuentes obtenidos de los CDs para extraer categorías subordinadas que consideramos hipónimos del hiperónimo. Para lograr lo anterior nos enfocamos en los adjetivos relacionales que comparten contenido semántico con el conjunto de hiperónimos. Asumimos que los adjetivos relacionales tienen una probabilidad mayor de ser buenos candidatos a hipónimos comparado con los adjetivos calificativos o descriptivos porque en lugar de un rasgo único vinculan al hiperónimo con un conjunto de rasgos o propiedades producto de su origen nominal.

5.4.1 Hiperónimo + modificador adjetivo

La tabla 28 muestra los hiperónimos con frecuencia mayor o igual que 9 obtenidos de los CDs del corpus de medicina y el total de candidatos que generan desglosados como relevantes y no relevantes al dominio en cuestión. Con este subconjunto se extraen todos los modificadores adjetivos del conjunto de FNs con estructura <NC><ADJ> y <NC><NC>, donde el hiperónimo es el núcleo nominal. Consideramos la estructura <NC>< NC > debido a que el porcentaje de error de

etiquetado más alto es precisamente un adjetivo que se etiqueta erróneamente como nombre común (NC), por lo que resulta de interés explorarlo. La precisión promedio de la extracción inicial de estas categorías más específicas, sin filtro alguno, es del 58% y la consideramos como nuestro *baseline*.

Tabla 28. Candidatos a hipónimos para hiperónimos con frecuencia de 9 o mayor.

Hiperónimo	Candidatos relevantes[23]	Candidatos no relevantes	Total de candidatos
Enfermedad	148	95	243
Trastorno	78	20	98
Examen	54	50	104
Afección	65	41	106
Procedimiento	19	29	48
Infección	98	76	174
Proteína	37	18	55
Cáncer	36	30	66
Tumor	42	25	67
Tratamiento	60	109	169
Cirugía	44	22	66
Método	28	47	75
Problema	76	47	123
Proceso	37	42	79
Inflamación	18	13	31
Glándula	20	21	41
Órgano	17	23	40
Medicamento	53	35	88
Total	930	743	1673

5.4.2 Filtrado de hipónimos no relevantes mediante información mutua estandarizada

La medida de información mutua se usa comúnmente para extraer colocaciones de fuentes textuales. De entrada, asumimos que esta medida puede ser de utilidad para medir la relación entre un hiperónimo y adjetivos relacionales con los que comparte contenido semántico, concretamente aplicamos la variante estandarizada propuesta por Bouma (2009), por los argumentos dados en la sección 4.5.3.

Para explorar la obtención de los mejores candidatos establecemos arbitrariamente varios umbrales de información mutua y analizamos los resultados en términos de los indicadores tradicionales: precisión, cobertura y medida F. La tabla 29 muestra los resultados obtenidos.

[23] Para determinar los candidatos relevantes nos apoyamos en la misma fuente MedLinePlus en español y en nuestra competencia lingüística.

Los resultados revelan poca utilidad de esta medida para extraer los hipónimos relevantes debido a que la cobertura se ve afectada significativamente a medida que aumentamos el umbral (C_PMI1>0, C_PMI2>0.10, C_PMI3>0.15, C_PMI4>0.25) y el efecto incremental en la precisión respecto al *baseline* (PSF) no es importante, como lo muestran las figuras 20 y 21, respectivamente. De los datos es claro que a medida que aumentamos el umbral perdemos información valiosa y no se logra un avance significativo en precisión. Dado lo anterior, suponemos que existe cierto grado de composicionalidad entre el hiperónimo y los modificadores relacionales, por tanto, exploramos heurísticas basadas en información lingüística para filtrar la información no relevante. El supuesto que guía nuestra búsqueda es la posibilidad de discernir, a nivel lingüístico, entre adjetivos calificativos o valorativos a un contexto específico, de aquellos relacionales que se derivan de un nombre y llevan implícitas todas las propiedades de dicho nombre. La siguiente sección da cuenta de estas heurísticas y cómo logran mejorar los resultados obtenidos con la medida de información mutua estandarizada.

Tabla 29. Evaluación de los candidatos a hipónimos.

Hiperónimo	**Precisión inicial**	**PMI1>=0**			**PMI2>=0.10**			**PMI3>=0.15**			**PMI4>=0.25**		
		P	**C**	**F**	**P**	**C**	**F**	**P**	**C**	**F**	**P**	**C**	**F**
Enfermedad	61	66	89	76	69	74	71	68	61	64	74	30	43
Trastorno	80	82	92	87	85	79	82	85	74	79	87	58	70
Examen	52	55	96	70	62	85	72	67	78	72	76	69	72
Afección	61	62	97	76	62	86	72	66	78	72	69	55	61
Procedimiento	40	40	100	57	43	100	60	40	84	54	57	68	62
Infección	56	58	92	71	64	81	72	67	67	67	73	47	57
Proteína	67	67	100	80	71	100	83	74	100	85	83	89	86
Cáncer	55	59	97	73	56	81	66	57	72	64	61	61	61
Tumor	63	63	100	77	64	88	74	66	83	74	69	64	66
Tratamiento	36	36	90	51	39	78	52	45	73	56	52	47	49
Cirugía	67	68	100	81	73	91	81	74	84	79	72	52	60
Método	37	37	100	54	39	100	56	39	93	55	38	64	48
Problema	62	64	93	76	63	75	68	65	54	59	67	29	40
Proceso	47	47	100	64	50	95	66	50	86	63	53	65	58
Inflamación	58	57	94	71	55	89	68	57	89	69	52	61	56
Glándula	95	95	100	97	95	100	97	95	100	97	95	95	95
Órgano	43	44	100	61	40	82	54	39	71	50	43	71	54
Medicamento	60	60	98	74	69	94	80	67	92	78	77	81	79
Promedio:	58	59	97	72	61	88	71	62	80	69	67	61	62

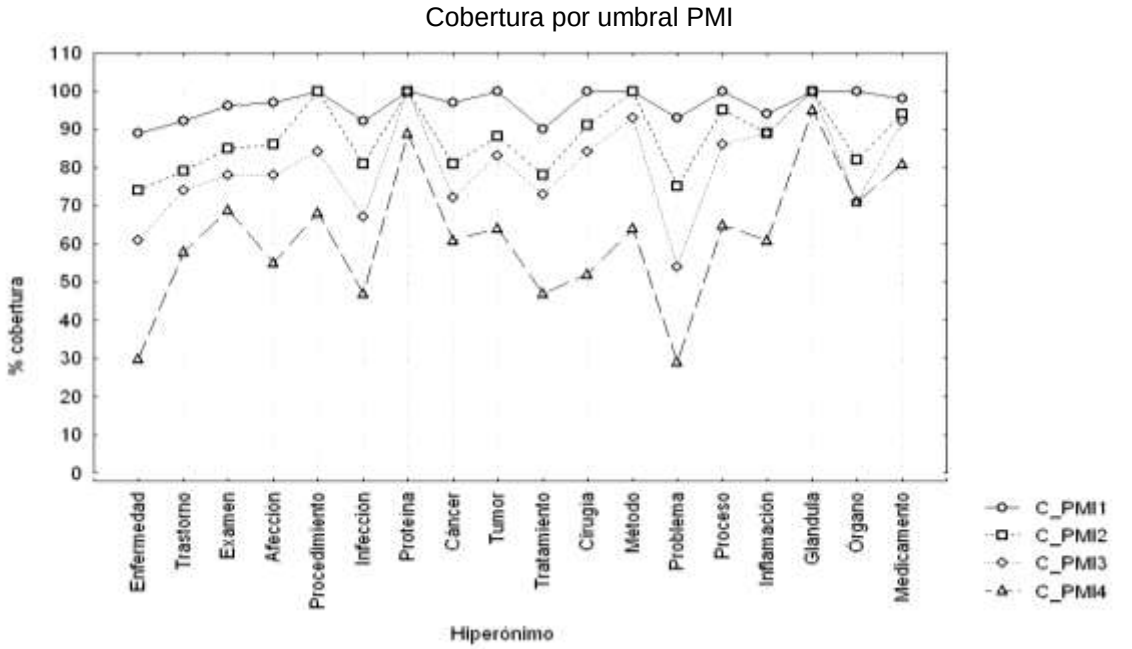

Figura 20. Cobertura por umbral PMI.

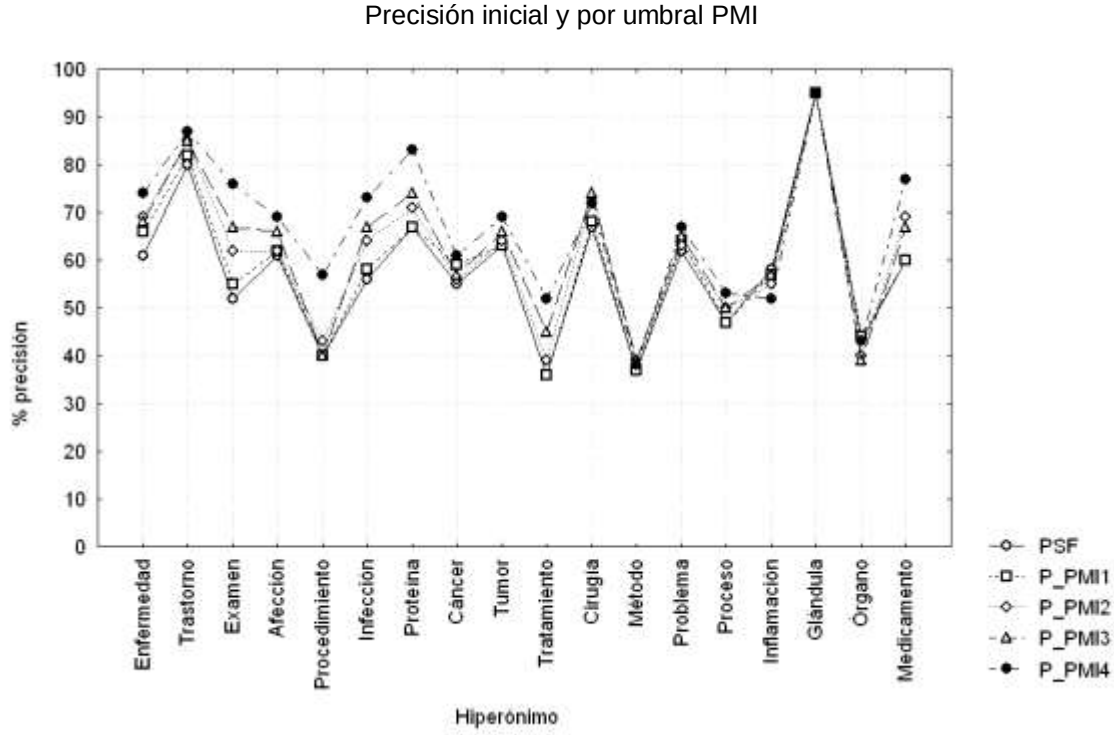

Figura 21. Precisión por umbral PMI.

5.4.3 Filtrado de adjetivos calificativos o descriptivos

Como se mencionó en la sección 4.5.4, exploramos heurísticas lingüísticas relacionadas con el comportamiento de adjetivos. Los adjetivos son una categoría abierta y no resulta práctico definir una lista de paro fija de adjetivos no relevantes de forma manual para filtrarse del conjunto de FNs previo a la extracción de los adjetivos relacionales que comparten contenido semántico con los hiperónimos. Dado lo anterior, proponemos la extracción de dichos adjetivos no relevantes del corpus de análisis aplicando dos de las heurísticas propuestas por Demonte (1999) para distinguir adjetivos calificativos de relacionales: graduabilidad y predicabilidad. Adicionalmente, incluimos otra más que considera la precedencia del adjetivo respecto al nombre que modifica. Dichas heurísticas derivaron en los resultados que se muestran en la tabla 30 y en la figura 22. De los datos de la tabla 30 resulta claro que existe una mejora significativa en cuanto a precisión (PSF=precisión sin filtro, PCF=precisión con filtro), con niveles aceptables de cobertura (CCF=cobertura con filtro) y medida F (FCF=F con filtro).

Tabla 30. Evaluación de heurísticas lingüísticas.

Hiperónimo	% Precisión Sin filtro (PSF)	Heurísticas lingüísticas		
		% PCF	% CCF	% FCF
Enfermedad	61	79	74	76
Trastorno	80	95	69	80
Examen	52	74	85	79
Afección	61	88	75	81
Procedimiento	40	85	89	87
Infección	56	84	76	80
Proteína	67	88	76	82
Cáncer	55	69	94	80
Tumor	63	82	86	84
Tratamiento	36	68	70	69
Cirugía	67	90	82	86
Método	37	72	82	77
Problema	62	83	67	74
Proceso	47	73	73	73
Inflamación	58	82	78	80
Glándula	95	100	100	100
Órgano	43	71	71	71
Medicamento	60	76	72	74
Promedio:	58	81	79	80

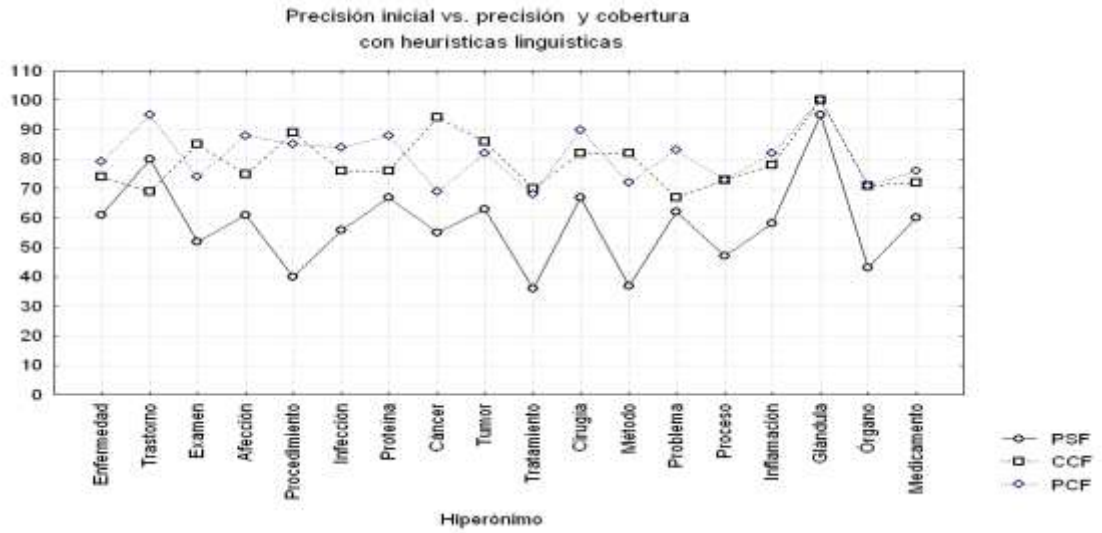

Figura 22. Desempeño de heurísticas lingüísticas.

Si comparamos el desempeño en cobertura con la aplicación de heurísticas lingüísticas (CCF) y el filtrado estableciendo umbrales PMI (C_PMI1>0, C_PMI2>.10, C_PMI3>.15, C_PMI4>.25), tenemos que con las heurísticas lingüísticas se mantiene por arriba de la cobertura lograda con un umbral de PMI>=0.25, excepto en dos casos, donde la cobertura es menor, como lo indica la figura 23.

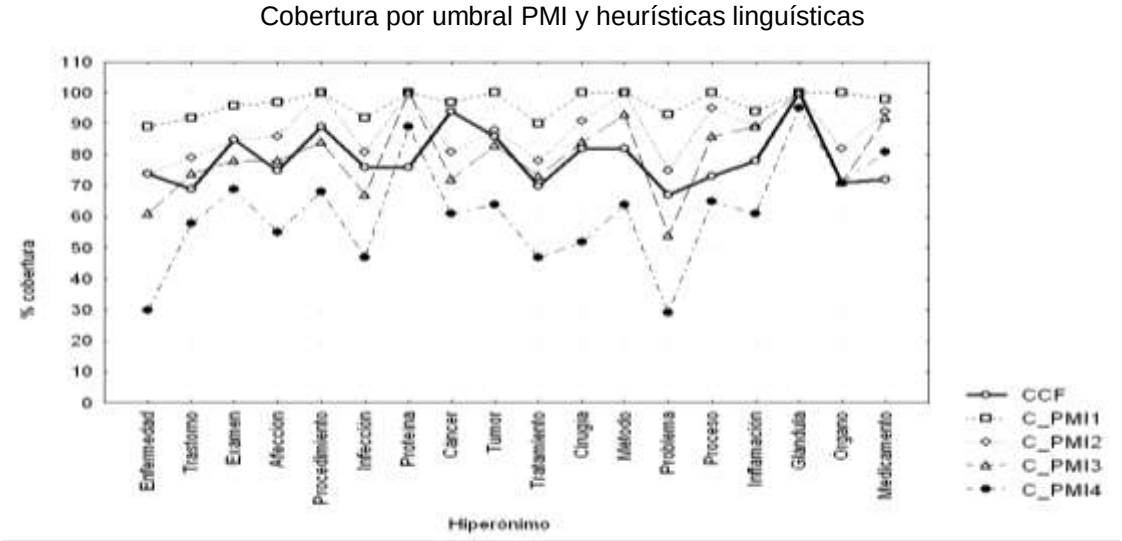

Figura 23. Desempeño considerando el nivel de cobertura.

Resultados de la extracción de relaciones de hiponimia-hiperonimia implícitas en CDs y en los hiperónimos más frecuentes

La precisión que se logra con la aplicación de nuestro método al corpus de medicina es de un 76%[24]. Esta precisión se logra considerando las relaciones extraídas de los CDs analíticos y de los hipónimos derivados de los hiperónimos más frecuentes listados en la tabla 28. Por otro lado, la cobertura obtenida es de un 12% de las relaciones implícitas en el corpus analizado.

5.4.4 Evaluación de la extracción de adjetivos no relevantes

La evaluación del proceso de extracción de adjetivos no relevantes por las tres heurísticas aplicadas produjo una cobertura del 45% de adjetivos que podrían no participar en la construcción de hipónimos relevantes o categorías más específicas y una precisión del 68% (ver tabla 31). Aunque la cobertura es baja, la aplicación de este subconjunto para filtrar relaciones no relevantes proporciona mejores resultados que aplicar umbrales de PMI.

Tabla 31. Evaluación del proceso de extracción de adjetivos no relevantes.

Adjetivos no relevantes	Adjetivos extraídos	Precisión	Cobertura	F
1946	866	68%	45%	54%

5.5 Un método para la extracción de hipónimos

Como se mencionó en la sección 1.2.2, el costo en tiempo y esfuerzo que implica la obtención manual de patrones e instancias de relaciones léxicas a partir de textos motivó el desarrollo de métodos que los *aprendieran* automáticamente. Uno de estos métodos probados para el español, pero independientes del lenguaje, es el propuesto por Ortega *et al.* (2007). Este método consiste en las tres etapas siguientes:

1. **Descubrimiento de patrones**. Extracción de un conjunto de patrones léxicos característicos de la relación de hiponimia utilizando un conjunto de instancias *semilla* de la relación para *aprender* de la Web. Posteriormente, se aplica un método de minería de textos sobre los

[24] Este resultado se obtiene promediando la precisión de la extracción conceptual y la precisión considerando las heurísticas lingüísticas.

patrones para obtener las secuencias de palabras maximales más frecuentes[25]. Dichas secuencias expresan los patrones léxicos que están más altamente relacionados con la relación de hiponimia. Finalmente, se seleccionan los patrones que satisfacen las siguientes expresiones regulares:

<cadena-frontera-izquierda> HIPÓNIMO <cadena-central>HIPERÓNIMO
HIPERÓNIMO<cadena-central>HIPÓNIMO<cadena-frontera-derecha>

2. **Extracción de instancias**. En esta fase, los patrones descubiertos en la primera etapa se aplican a un conjunto de documentos para localizar fragmentos de texto que sean candidatos a contener una instancia de la relación de hiponimia. El resultado de esta fase, entonces, es un conjunto de instancias candidatas a pares hipónimo-hiperónimo.

   Una de las particularidades del método es que basados en el vocabulario proporcionado por el usuario construye consultas fijando el hiperónimo, por ejemplo, si se tiene el patrón "El HIPÓNIMO es uno de los HIPERÓNIMOS" y el hiperónimo *piedra*, el método construye la consulta "el HIPÓNIMO es una de las piedras". Así, con esta funcionalidad es posible extraer más hipónimos del concepto *piedra.*
3. **Clasificacion de instancias**. El objetivo de este módulo es evaluar la confianza de las instancias extraídas, de tal suerte que las que tengan una probabilidad más alta de ser verdaderas se encuentren en las primeras posiciones de la lista de resultados. *Grosso modo*, el criterio de evaluación consiste en que las instancias más pertinentes se extraen por diferentes patrones, y a su vez los patrones valiosos extraen varias instancias pertinentes. Para lograr lo anterior, se definió un proceso de evaluación iterativo donde la confianza de las instancias se calcula con base en la confianza de los patrones, y viceversa.

Aplicamos el método propuesto por Ortega *et al.* (2007) al corpus de medicina utilizando un conjunto de instancias *semilla* de la relación de hiponimia-hiperonimia obtenidos automáticamente de los CDs con la metodología

[25] Una secuencia de palabras maximal frecuente es una cadena de palabras cuya ocurrencia excede un umbral especificado de antemano y que no es una subsecuencia de otra secuencia frecuente (Ortega *et al.*, 2007).

propuesta en este trabajo. El conjunto de instancias corresponde a los CDs que contienen los hiperónimos más frecuentes y cuyo término es unipalabra (111 pares de la relación). De este subconjunto, seleccionamos aleatoriamente pares de hipónimo-hiperónimo. El objetivo de la aplicación de este método a nuestro corpus es determinar su pertinencia para la extracción de relaciones léxicas de hiponimia-hiperonimia en colecciones cerradas aprovechando un conjunto de instancias obtenidas de los CDs. Las instancias *semillas* se presentan en la tabla 32.

Tabla 32. Conjunto de instancias *semilla*.

Hipónimo	**Hiperónimo**
Coriocarcinoma	Cáncer
Alcaptonuria	Trastorno
Angioplastia	Procedimiento
Artrosis	Enfermedad
Autismo	Trastorno
Colonoscopía	Procedimiento
Coriocarcinoma	Enfermedad
Craneosinostosis	Afección
Digitálico	Medicamento
Muermo	Enfermedad
Obesidad	Enfermedad
Páncreas	Órgano
Prostatitis	Inflamación
Regionalización	Procedimiento
Rubéola	Enfermedad
TB	Enfermedad
TDAH	Trastorno
Troponina	Proteína
Electronistagmografía	Examen
Endoftalmitis	Afección
Fibrodisplasia	Enfermedad
Gastroparesia	Enfermedad
Glucagonoma	Tumor
Pericarditis	Afección
Sigmoidoscopía	Procedimiento

Resultados

La tabla 33 muestra el número de ejemplos localizados en el corpus para las *semillas* mostradas en la tabla 32. Los ejemplos se utilizan para *aprender* los patrones más característicos de la relación de hiponimia.

Tabla 33. Número de ejemplos por flexión de número.

Flexión hipónimo-hiperónimo	Número de ejemplos
Singular-singular	75
Singular-plural	25
Plural-plural	7
Plural-singular	0

Con un umbral de frecuencia de patrones β=2, se obtuvieron los patrones mostrados en la tabla 34 que cumplen con los requerimientos del método.

Tabla 34. Conjunto de patrones.

Flexión hipónimo-hiperónimo	Patrones
Singular-singular	[3][26] la,<HIJO>,es,un,<PADRE> [2] el,<HIJO>,es,un,<PADRE> [2] la,<HIJO>,es,una,<PADRE>, [2] la,<HIJO>,aumenta,el,riesgo,de,<PADRE> [2] la,<HIJO>,sea,clasificada,como,una,<PADRE> [2] la,<HIJO>,constituye,la,<PADRE>,
Singular-plural	[2] <PADRE>,asociadas,con,la,<HIJO>,y [2] la,<HIJO>,es,una,de,las,primeras,<PADRE>, [2] la,<HIJO>,se,convierte,en,una,de,las,<PADRE>, reemergentes

Posteriormente, se generaron las expresiones regulares de los patrones para realizar la búsqueda de nuevas instancias en el corpus de medicina. Las expresiones regulares se muestran en la tabla 35. No se localizaron ocurrencias con la flexión plural-plural y plural-singular, por ello no aparecen en la tabla 35. La expresión hiperónimo_i representa el vocabulario que debe corresponderse en el corpus. La lista es la siguiente:

(inflamación | proteína | trastorno | examen | enfermedad | órgano | cirugía | procedimiento | medicamento | afección)

Tabla 35. Expresiones regulares asociadas con los patrones.

Flexión hipónimo-hiperónimo	**Expresiones regulares**	**Relevantes /extraídos**
Singular-singular	la\s+([\w\s+]+)\s+es\s+un\s+(hiperónimo$_i$) el\s+([\w\s+]+)\s+es\s+un\s+(hiperónimo$_i$) la\s+([\w\s+]+)\s+es\s+una\s+(hiperónimo$_i$) la\s+([\w\s+]+)\s+aumenta\s+el\s+riesgo\s+de\s+(hiperónimo$_i$) la\s+([\w\s+]+)\s+sea\s+clasificada\s+como\s+una\s+(hiperónimo$_i$) la\s+([\w\s+]+)\s+constituye\s+la\s+(hiperónimo)	2/6 3/11 6/10 1/1 2/2 0/1
Singular-plural	(hiperonimo)\s+asociadas\s+con\s+la\s+([\w\s+]+) la\s+([\w\s+]+)\s+es\s+una\s+de\s+las\s+primeras\s+(hiperónimo$_i$) la\s+([\w\s+]+)\s+se\s+convierte\s+en\s+una\s+de\s+las\s+(hiperónimo$_i$)	2/2 1/1 1/1

[26] El número indica la frecuencia de ocurrencia del patrón.

Búsqueda de instancias nuevas en corpus

La búsqueda de nuevas instancias en el corpus generó un conjunto de 35 candidatos, de los cuales sólo 18 corresponden a relaciones de hiponimia-hiperonimia verdaderas, lo cual genera una precisión del 51% y una cobertura muy baja del 0.15%.

Conclusión del proceso

De los resultados obtenidos con el método propuesto por Ortega *et al.* (2007) podemos concluir lo siguiente:

1. Presenta limitaciones para colecciones cerradas debido a que la probabilidad de obtener una gran variedad de ejemplos para *aprender* los patrones es baja. Del mismo modo, la evaluación de la confiabilidad de los patrones e instancias se dificulta debido a la escasez de datos. Creemos que las colecciones cerradas deben tener un tamaño mucho mayor y relacionarse con una temática específica para obtener mejores resultados.
2. La eliminación de la puntuación en los patrones impacta negativamente la extracción de nuevas instancias en colecciones cerradas porque muchos patrones eficaces se expresan mejor usando una combinación de palabras y signos de puntuación.
3. Los resultados obtenidos en colecciones cerradas no son directamente comparables con los obtenidos por los autores del método debido a que éste último se basa en un conjunto grande de *snippets* para cada instancia hipónimo-hiperónimo *semilla*, lo que aumenta la probabilidad de obtención de distintas formas para expresar la misma relación y posibilita la medición de confiabilidad de instancias y patrones.
4. Consideramos que este tipo de métodos deben aplicarse una vez en grandes conjuntos de textos para *aprender* y evaluar los patrones característicos de una relación. Una vez aprendidos y determinada su confiabilidad pueden aplicarse en colecciones cerradas.
5. Respecto a la selección de las instancias *semillas* para aprender los patrones consideramos que puede influir un efecto de *tipicalidad* de los términos incluidos. Es decir, entre más característico o común sea el par hipónimo-hiperónimo en el dominio, más probable será que se exprese en

una gran variedad de contextos.

6. A pesar de que en todas las metodologías sobre extracción conceptual y relaciones léxicas de hiponimia-hiperonimia se habla de la gran variedad de contextos en los que se usa el verbo "ser" y de la cantidad de ruido que introduce en los resultados, sigue siendo el más utilizado para expresar la relación. Dada esta situación, las medidas de confiabilidad adoptadas por Ortega *et al.* (2007) resultan muy pertinentes para filtrar la información relevante a partir de considerar que si una instancia se expresa con más de un patrón característico de hiponimia, entonces su probabilidad de ser relevante es alta. De igual forma, si un patrón presenta varias instancias relevantes, entonces es un buen patrón.

5.6 Discusión de resultados

En este capítulo hemos dado cuenta de la implementación de la metodología propuesta en este trabajo de investigación. Uno de los puntos de partida de nuestra metodología es la extracción de información conceptual mediante la consideración de un filtro sintáctico que contemple la estructura más común de términos en dominios especializados y aquellas donde regularmente podemos encontrar un hiperónimo, ambos enmarcados por un patrón verbal específico que permite su localización y extracción. Las condiciones sintácticas especificadas *a priori* reducen nuestra cobertura comparado con la obtenida por la herramienta Ecode, sin embargo, se logra reducir significativamente el ruido, lo que se ve reflejado en una mayor precisión.

Como se mencionó en la sección 4.3.2, el comportamiento sintáctico de los CDs no es suficiente para determinar si un fragmento es un buen candidato o no. Es indispensable la implementación de otras estrategias para eliminar el *ruido* sin impactar de forma dramática la cobertura. En este trabajo implementamos un filtro semántico que tenía como objetivo eliminar los fragmentos de CDs con núcleos nominales indicadores de otro tipo de relación, sobre todo haciendo énfasis en el verbo *ser*, que es el que se utiliza en una mayor cantidad de contextos y da origen a una gran cantidad de *ruido*. Las relaciones que enfatizamos fueron las de parte-todo y causalidad debido a que han sido de las más trabajadas en la literatura sobre la extracción automática de relaciones a partir de información textual. Como resultado de la aplicación del filtro semántico

observamos un aumento en precisión y un impacto prácticamente nulo en cobertura, lo cual nos da elementos para sugerir su pertinencia previo a la tipificación de otras relaciones que también se expresen con dicho núcleo verbal.

En términos generales, de acuerdo con los resultados obtenidos, por lo menos un 50% de los CDs tiene una estructura más cercana a lo canónico, que es la que logramos capturar con nuestros patrones. El resto que no se logró capturar es una combinación de CDs con estructura más compleja, así como también de limitaciones del etiquetado de partes de la oración que impiden la extracción de éstos mediante los patrones considerados.

Un aporte importante de nuestro trabajo es la exploración de los hiperónimos para obtener categorías más específicas, en el sentido descrito en la sección 2.4.4. Como se discutió en dicha sección, investigaciones realizadas en psicología cognitiva han mostrado resultados relacionados con el uso de un nivel básico. Algunos resultados evidenciaron que este nivel básico podría ser el *genus*, sin embargo, resultados posteriores mostraron que el nivel básico podría variar de acuerdo con el nivel de entrenamiento de la persona realizando la categorización, por lo que se propuso que para personas con un mayor entrenamiento en un dominio determinado, el nivel básico podría ser una categoría subordinada al *genus*.

Cruse (2002) señala que un buen hipónimo de un hiperónimo no necesariamente es un taxónimo. Para ser un taxónimo debe reflejar una perspectiva de división relevante o útil para el dominio. De acuerdo con este mismo autor, generalmente estos casos se reflejan mediante la descripción del hipónimo vía el hiperónimo más un simple rasgo, por ejemplo, *semental = caballo macho*. Contrario a la situación anterior, los adjetivos relacionales, dado su origen nominal, vinculan al nombre que modifican con un conjunto de rasgos, por ello consideramos son relevantes como relaciones de hiponimia o categorías subordinadas al *genus*.

Finalmente, de acuerdo con los resultados obtenidos con nuestro método, logramos una mejor precisión y cobertura a nivel global considerando todas las relaciones de hiponimia-hiperonimia implícitas en el corpus de medicina que los métodos de extracción conceptual y de relaciones de hiponimia que se probaron en esta investigación. Como se mencionó anteriormente, los métodos de

aprendizaje de patrones son efectivos en grandes colecciones de textos que garanticen la presencia de una gran variedad de patrones característicos de la relación y consideramos que el corpus de medicina es una colección cerrada donde enfrentamos el problema de escasez de datos.

Capítulo 6

Conclusiones y trabajo futuro

6.1 Conclusiones

6.1.1 Enfoque clásico y definiciones analíticas

La metodología que se propuso en este trabajo se basa en patrones por dos razones principales. En primera instancia, los resultados de investigaciones respecto a la extracción conceptual de textos en español indican que un enfoque basado en patrones léxico-sintácticos es factible dada la regularidad que presentan las definiciones en textos técnicos y científicos, además de que este comportamiento no se circunscribe a un dominio específico. Por otro lado, dado que usamos un análisis sintáctico superficial (chunking), en lugar de uno completo para reconocer la información relevante, consideramos que nuestro método es escalable a grandes colecciones de textos. Sin embargo, una desventaja importante es la dependencia del lenguaje.

La regularidad sintáctica en la formulación de conceptos se debe en buena parte a la adopción del enfoque clásico como una meta-teoría en dominios especializados. Sin embargo, a pesar de su uso común, esta perspectiva clásica no ha estado exenta de controversias. Entre las principales críticas se encuentran la imposibilidad de definir una gran cantidad de conceptos en términos de propiedades necesarias y suficientes, así como también de los efectos de *tipicalidad* de algunos elementos respecto a otros de la misma clase.

Las definiciones analíticas representan el enfoque prototípico de la visión clásica para enunciar conceptos. Es justo en estos esquemas de definición donde se evidencian los problemas con el enfoque clásico, algunos de los cuales se relacionan con la asignación de más de un *genus* al concepto definido en diferentes CDs. A partir del comportamiento observado en el conjunto de CDs extraído en la implementación de nuestra metodología, una de las aportaciones de nuestro trabajo se relaciona precisamente con la explicación del origen de estas diferentes categorías. *Grosso modo*, estas categorías pueden evidenciar, en primera instancia, la falta de consenso respecto a la clase a la que pertenece el concepto definido. Por otro lado, dichas clases pueden corresponder a algún nivel

dentro de una estructura jerárquica conceptual, siendo las más comúnmente utilizadas en dominios especializados aquellas que proporcionen mayor información respecto al término definido, como son el *genus* o categorías subordinadas. Otro caso más lo constituyen las variantes léxicas o sinónimos de una misma categoría. Finalmente, dichas categorías pueden reflejar un rol funcional o agentivo manifiesto lingüísticamente en la forma de una nominalización verbal donde el término definido es parte del conjunto de argumentos del verbo.

6.1.2 Eliminación de ruido

El *ruido* presente en los resultados es un problema muy común en los procesos de extracción de información. En nuestro caso, dado que los verbos considerados no sólo se utilizan en CDs obtuvimos un porcentaje alto de *ruido*, sobre todo con el verbo *ser*, que es el más polisémico. Para enfrentar este problema, sin afectar dramáticamente la cobertura, se implementaron tres filtros. El primer filtro se enfocó en la sintáxis de los elementos más relevantes: término, patrón verbal e hiperónimo, así como también de otros constituyentes tales como patrones pragmáticos y sinónimos del término. La consideración de la estructura sintáctica de un CD hace dependiente el método a un lenguaje específico, sin embargo, es proveedora de significado y sirve como filtro para extraer el comportamiento más canónico de los CDs. Con el filtro sintáctico logramos una precisión del 60%, que está por encima del 24% obtenido con la herramienta de extracción conceptual Ecode, sin embargo, la cobertura es menor debido a las restricciones sintácticas de los constituyentes considerados de forma *a priori*. Por otro lado, el filtro semántico permite filtrar otro tipo de relaciones que no son de interés. En términos generales, con el filtro semántico logramos un aumento en precisión del 68% y con efectos nulos en cobertura, lo que constituye un resultado importante y deja abierta la posibilidad de tipificar otro tipo de relaciones para filtrar una mayor cantidad de información no relevante.

Un tercer filtro considera el conjunto de *genus* o hiperónimos extraídos para filtrar el conjunto de CDs candidatos. El argumento que subyace en la aplicación de este filtro es que los hiperónimos más frecuentes son los que tienen una probabilidad mayor de ser verdaderos, sin embargo, para lograr mejores resultados, la colección de textos debe ser grande y relacionada con un dominio

específico. Consideramos que el subconjunto de relaciones de hiponimia-hiperonimia extraídos de los hiperónimos más frecuentes tiene una confiabilidad alta y puede ser usado como conjunto *semilla* en procesos de aprendizaje automático de patrones e instancias. A este respecto, la implementación del método propuesto por Ortega *et al.*, (2007) con las instancias *semilla* de los hiperónimos más frecuentes derivó en una precisión del 51%, sin embargo, la cobertura es muy baja (0.15%), lo que está por debajo del desempeño alcanzado con la aplicación de nuestra metodología. Las causas de lo anterior pueden ser varias, sin embargo, el problema principal es la escasez de datos en colecciones cerradas que no garantiza la existencia de varias formas de expresión de la relación, lo cual influye directamente en los mecanismos de evaluación de patrones e instancias.

6.1.3 Hipónimos de un hiperónimo

Los hipónimos o categorías subordinadas al hiperónimo pueden ser buenos taxónimos. Para obtener los mejores candidatos, exploramos una medida como la información mutua para determinar la relación entre un núcleo nominal (hiperónimo) y modificadores, tales como adjetivos y nombres comunes. Para el caso de los adjetivos, nos enfocamos en los relacionales porque son conceptualmente complejos y vinculan al hiperónimo con un conjunto de propiedades derivadas de su origen nominal, por lo que pueden ser de utilidad como categorías subordinadas. Entre los resultados obtenidos encontramos que existe una buena cantidad de adjetivos relacionales que son relevantes para los dominios en cuestión y que constituyen perspectivas de división del hiperónimo interesantes. Por otro lado, los nombres comunes son en gran parte adjetivos que fueron etiquetados erróneamente. Como se recordará del análisis de etiquetado de partes de la oración presentado en la sección 4.2.5, aproximadamente un 30% de los errores son de este tipo.

Dada la composicionalidad que guardan los adjetivos relacionales con los hiperónimos, la medida de información mutua no produjo buenos resultados para filtrar información no relevante. Debido a esta situación, probamos tres heurísticas lingüísticas para discernir adjetivos calificativos de relacionales. Las heurísticas explotan la regularidad del lenguaje para obtener una lista de paro de adjetivos, en su gran mayoría calificativos o descriptivos, de tal suerte que

puedan ser filtrados antes de ofrecer los resultados finales. Dichas heurísticas lograron mejores resultados que el establecimiento de umbrales de información mutua. La implementación computacional de estas heurísticas vía un análisis sintático superficial constituye otro aporte importante de nuestro trabajo.

6.1.4 Extracción de información

La meta más ambiciosa que se pretende alcanzar con el desarrollo de nuevas aplicaciones de extracción de información es limitar o, si fuese posible, eliminar la necesidad de intervención humana en el proceso de adaptación y portabilidad de aplicaciones de extracción de información a nuevos dominios. Esto cobra mayor importancia si nos planteamos la función de un agente inteligente que explore la Web, donde hay una gran cantidad de fuentes de información heterogéneas, en la búsqueda de información relevante para cubrir las necesidades cada vez más demandantes de sus usuarios.

Retomando el caso de la extracción automática de relaciones léxico-semánticas se han explotado básicamente tres enfoques: 1) patrones, 2) información distribucional y 3) subsunción de conceptos. Es notable todavía la consideración de patrones característicos de la relación extraídos manualmente o *aprendidos* de corpus. Los avances más importantes respecto a la propuesta inicial de Hearst son la incorporación de medidas de confiabilidad para rankear las relaciones obtenidas y la aplicación de aprendizaje supervisado. Este último todavía resulta muy costoso debido a que la creación de datos de entrenamiento apropiados para el proceso de aprendizaje no es una tarea trivial y requiere de la participación de expertos.

Finalmente, consideramos que los dominios restringidos siguen siendo los casos más exitosos para la extracción de información debido a que se encuentran, en algún punto del tiempo, circunscritos a un conjunto de conceptos y con ello contribuyen a filtrar la información relevante. En el caso particular de la metodología que proponemos, los hiperónimos más frecuentes aumentan la precisión de la extracción y contribuyen también a reflejar la temática de la colección analizada.

6.1.5 Extracción de relaciones léxico-semánticas

Mucho se ha hablado en la literatura de extracción de información sobre un tema concreto: extracción de relaciones léxicas o semánticas, o bien, léxico-semánticas. Para empezar, existe un uso indistinto de los términos léxico y semántico para denotar las relaciones que son de interés. Esta falta de consenso no tiene su origen en la comunidad computacional, sin embargo, se promueve. El origen de esta controversia radica en la discusión de si las relaciones realmente se dan entre las palabras, o los sentidos de éstas, dado que una misma palabra puede tener más de un significado, o bien, entre los sentidos de las palabras en contexto. A la fecha no es claro todavía que término es el más adecuado para etiquetar este tipo de relaciones, sin embargo, existe ya un buen camino recorrido en su identificación a partir de textos.

En esta investigación decidimos aprovechar el caudal de resultados derivados de áreas como la lingüística computacional y cognitiva, terminología, y psicología cognitiva para proponer una metodología de extracción de relaciones léxico-semánticas de hiponimia-hiperonimia confiable que sirva como conjunto *semilla* en procesos de aprendizaje automático. Sin duda, nuestro énfasis principal fue la extracción: cuánto recuperamos y con qué precisión, porque son indicadores, por lo menos desde una perspectiva computacional, de gran relevancia para efectos de comparar nuevos enfoques con los ya existentes y medir con ello el desempeño. En este sentido, para el caso de colecciones cerradas, logramos una precisión del 76% y una cobertura del 12%, que está por encima del desempeño de los métodos de extracción de relaciones de hiponimia-hiperonimia propuestos a la fecha para el español.

Aunado a lo anterior, nuestro trabajo se ocupa también de dar cuenta de los problemas que se pueden encontrar cuando se extrae este tipo de relación de definiciones en textos e intenta explicar su origen, así como también sus repercusiones para tareas de representación de conocimiento. Parte de estos problemas ya habían sido considerados, sin embargo, no existen aplicaciones concretas que los aborden.

En resumen, consideramos que los resultados de esta investigación culminan con un enfoque adecuado que logra resultados importantes porque constituye una vía para automatizar el proceso de selección de instancias semilla

para propósitos de *aprendizaje* de patrones y más instancias.

6.2 Trabajo futuro

6.2.1 Problemas encontrados

Una de las desventajas importantes de los enfoques léxico-sintácticos para la extracción de información es que son dependientes del lenguaje para el que se desarrollan debido a que requieren de un proceso de etiquetado de partes de la oración o sintáctico, de tal suerte que sea posible la definición de expresiones regulares o reglas para capturar el comportamiento de la información relevante. Por otro lado, una de las ventajas radica en el poder de generalización de dichos patrones. En el análisis del proceso de etiquetado de partes de la oración presentado en la sección 4.2.5 se obtuvo una precisión del 76% y se describieron los errores más comunes. Como se mencionó también en dicha sección, los etiquetadores de partes de la oración se entrenan generalmente con corpus de lengua general o de otro dominio diferente al que se intentan aplicar, lo que reduce su precisión. Para aumentar la precisión se han propuesto diversos métodos, sin embargo, al tiempo de realizar este trabajo no se encontraron implementaciones disponibles para probar su eficacia, por lo que consideramos es un campo abierto para nuevas propuestas.

Por otro lado, otra más de las grandes limitaciones para realizar este tipo de investigaciones es la falta de recursos textuales en español adecuados, en tipo y tamaño, para probar los diferentes enfoques propuestos. Por ejemplo, en este sentido, representaría una gran ventaja contar con corpus de otros dominios para efectos de realizar comparaciones sobre la estabilidad de los resultados obtenidos.

6.2.2 Retos de los métodos de extracción de relaciones de hiponimia-hiperonimia

1. Como se mencionó en apartados anteriores, actualmente se ponderan más los métodos adaptables y escalables a nuevos dominios, así como también a otros lenguajes. Por lo que resulta indispensable probar estos requerimientos en la metodología propuesta.

2. Los métodos propuestos a la fecha para la extracción automática de relaciones de hiponimia no hacen distinción alguna de los diferentes tipos que pueden encontrarse. Creemos que esto se debe más que nada a que dichos métodos priorizan la etapa de extracción y no la utilidad de estas relaciones, por ejemplo, para efectos de representación de conocimiento. Wierzbicka (1984), Pustejovsky (1991) y Fellbaum (1998) son los referentes más inmediatos en cuanto a los tipos de hiperonimia que se pueden encontrar asociados con un elemento léxico. Es indispensable realizar un análisis más profundo de esta tipología de relaciones de hiperonimia ya que, como bien lo señala Pustejovsky, revelan la polisemia de un elemento léxico y contribuyen a dar cuenta de su significado.

Bibliografía

Acosta, O., Sierra, G., Aguilar, C. 2011: "Extraction of Definitional Contexts using Lexical Relations". *International Journal of Computer Applications*. Volume 33(6): 46-53.

Acosta O., Aguilar C., Sierra G. 2010: "A Method for Extracting Hyponymy-Hypernymy Relations from Specialized Corpora Using Genus Terms". En Alemany, L. A., et al. (eds.), *Proceedings of the Workshop in Natural Language Processing and Web-based Technologies 2010*, Universidad Nacional de Córdoba, Córdoba, Argentina: 1-10.

Acosta, O. 2009: "Automatic Extraction of Lexical Relations from Analytical Definitions Using a Constrait Grammar". In *Advances in Artificial Intelligence.* 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009. Lecture Notes in Computer Science, Springer. Volume 5549/2009: 262-265.

Agirre, E., Soroa, A. 2007: "Evaluating Word Sense Induction and Discrimination Systems". In SemSeval-2007. Association for Computational Linguistics.

Aguilar, C. 2009: *Análisis lingüístico de definiciones en contextos definitorios*. Tesis doctoral. Posgrado en Lingüística-UNAM.

Aguilar C., Acosta O., Sierra G. 2010: "Recognition and Extraction of Definitional Contexts in Spanish for Sketching a Lexical Network". En Solorio, Th. & Pedersen, E. (eds.), *Proceedings of 1st Young Investigators Workshop on Computational Approaches to Languages of the Americas*, Association of Computational Linguistics Publications, Stroudsburg, PA.: 109-116.

Aguilar, C., Sierra, G., Acosta, O. 2012: "Thinking about a Conceptual Inquiry into Medical Texts: a Work in Progress", *Innovative Ways of Knowledge Representation and Management*, Universidad de Medellín, Colombia. En Prensa.

Sierra, G., Cruz, I., Acosta, O. 2010: "El sintagma nominal en la extracción de relaciones léxico-semánticas de Contextos Definitorios: el caso de la preposición *de"*. En prensa.

Alarcón, R. 2009: *Descripción y evaluación de un sistema basado en reglas para la Extracción automática de contextos definitorios*. Ph. D. Dissertation. IULA-UPF, Barcelona.

Allan, K. 1986. *Linguistic meaning*. London: Routledge. Natural language semantics. Oxford: Blackwell.

Alsawhi, H. 1987: "Processing dictionary definitions with phrasal pattern hierarchies". *Computational Linguistics* 13(3-4): 195-202.

Amsler, R. 1981: "A taxonomy for English nouns and verbs". In: *Proceedings 19th Annual Meeting of the Association for Computational Linguistics*: 133-38.

Amrani, A., Kodratoff, Y., Matte-Taillez, O. 2004: "A Semi-automatic System for Tagging Specialized Corpora". In Advances in Knowledge Discovery and Data Mining. *Proceedings Lecture Notes in Artificial Intelligence*. Subseries of Lecture Notes in Computer Science, 3056 (3056). Springer, Heidelberg, Germany: 670-681.

Baader, F., McGuinness, D., Nardi, D. 2003: *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.

Ballesteros, M., Herrera J., Francisco, V., Gervás, P. 2010: "Improving Parsing Accuracy for Spanish Using Maltparser". *Procesamiento de Lenguaje Natural*, Revista No. 44.

Barrón, A. 2007. *Extracción Automática de Términos en Contextos Definitorios*. Tesis de maestría. Posgrado en Ciencia e Ingeniería de la Computación-UNAM.

Barsalou, L., Simmons, W., Barbey, A., Wilson, C. 2003: "Grounding conceptual knowledge in modality-specific systems". *Trends in Cognitive Sciences*, 7: 84-91.

Berland, M., Charniak, E. 1999: "Finding parts in very large corpora". In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*: 57-64.

Berlin, B., Breedlove, D., Raven, P. 1974: *An Introduction to the Botanical Ethnography of a Mayan-Speaking People of Highland Chiapas*. Academic Press (New York).

Berlin, B. 1992. *Ethnobiological classification: principles of categorization of plants and animals in traditional societies.* Princeton, N.J.: Princeton University Press.

Bird, S., Klein, E., Loper, E. 2009: *Natural Language Processing whit Python*. O'Reilly, Sebastropol, CA.

Bouma, G.2009: "Normalized (Pointwise) Mutual Information in Collocation Extraction". In: Chiarcos, C., Castilho, E., Stede, M. (eds). From Form to Meaning: Processing Texts Automatically, *Proceedings of the Biennial GSCL Conference 2009*: 31-40.

Bourigault, D. 1994. *LEXTER, un Logiciel d'Extraction de TERminologie. Application à l'acquisition des connaissances à partir de texts*. PhD Thesis. Paris: Ecole des Hautes Etudes en Sciences Sociales, Paris.

Brachman, R., Levesque, H. 1985: *Readings in knowledge representation*. Morgan Kaufmann, Los Altos, CA.

Brown, C. 1984: *Language and Living Things: Uniformities in Folk Classification and Naming*. New Brunswick, Rutgers University Press.

Buitelaar, P., Cimiano, Ph., Magnini, B. 2005: *Ontology learning from text*. IOS Press, Amsterdam.

Buitelaar, P., Cimiano, P. 2008: *Ontology Learning and Populations: Bridging the Gap between Text and Knowledge*. Frontiers in Artificial Intelligence and Applications. IOS Press.

Cabré, M., Aldestein, A. 2000. *¿Es la terminología lingüística aplicada?.* Actas del XVIII Congreso de AESLA, Barcelona.

Calvo, F. 2006: *Determinación automatic de roles semánticos usando preferencias de selección sobre corpus muy grandes*. Tesis doctoral. Laboratorio de lenguaje natural-IPN.

Caraballo, S. 1999: "Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text". In *Proceedings of ACL '99*, College Park, MD.

Charniak, E., Berland, M. 1999: "Finding parts in very large corpora". In *Proceedings of the 37th.*

Cimiano, Ph., Pivk, A., Schmidt, L., Staab, S. 2004: "Learning Taxonomic Relations from Heterogeneous Sources of Evidence". In: *Proceedings of the ECAI 2004 Ontology Learning and Population*, Valencia, Spain.

Croft, W., Cruse, A. 2004: *Cognitive Linguistics*. Cambridge University Press.

Cruse, A. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Cruse, A. 2002. "Hyponymy and its varieties". In R. Green, C. A. Bean, and S. H. Myaeng (eds.), *The semantics of relationships: An interdisciplinary perspective*. Dordrecht: Kluwer: 3–22.

Daille, B. 2003. "Conceptual Structuring through Term Variations". Proc. *ACL Workshop on MultiWord expressions : Analysis, Acquisition and Treatme*nt.

Demonte, V. 1999. "El adjetivo. Clases y usos. La posición del adjetivo en el sintagma nominal". *En Gramática descriptiva de la lengua española*, Vol. 1, Cap. 3: 29-215.

De Saussure, F. (1916/1996). *Cours de linguistique générale.* Paris: Payot.

Dorr, B., Garman, J., Weinberg, A. 1994: "From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT". IN *Machine Translation,* Volume 9: 221-250

Drouin, P. 2003. "Term Extraction Using Non-Technical Corpora as a Point of Leverage". *Terminology*, Vol. 9, Number 1, John Benjamins Publishing Company: 99-115.

Enguehard, C. 1993. "Acquisition de Terminologie à partir de Gros Corpus". *Informatique & Langue Naturelle*, ILN'93, Nantes: 373-384.

Estopà, R.2003: *Extracció de terminologia: elements per a la construcció d'un* SEACUSE. Ph. D.Dissertation, IULA-UPF, Barcelona.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Soderland, S., Popescu, A., Weld, D., Shaked, T., Yates, A. 2004: "Web-Scale Information Extraction in KnowItAll (Preliminary Results)". In *Proceedings of the 13th World Wide Web Conference*: 100–109.

Evans, V., Green, M. 2006: *Cognitive Linguistics: An introduction*. Mahwah, NJ: Lawrence Erlbaum Associates.

Evans, R. 2003: “A framework for Named Entity Recognition in the Open Domain”. In *Proceedings of the Recent Advances in Natural Language Processing* (RANLP-2003):137-144.

Faure, D., Nédellec, C. 1998: “A Corpus-based Conceptual Clustering Method for Verb Frames and Ontologies”. *Proceedings of the LREC Workshop on adapting lexical and corpus resources to sublanguages and applications*: 5-12.

Fellbaum, C. 1998: *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press: 1–19.

Fellbaum, C. 2005. “WordNet and WordNets”. In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier: 665-670.

Fillmore, Ch. 1975: “An Alternative to Checklist Theories of Meaning”. *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society,* ed. Cathy Cogen et al.: 123-31.

Garduño, G., Sierra, G., Medina, A. 2004. “Herramientas de análisis para el Corpus Lingüístico en Ingeniería”. En *Avances en la Ciencia de la Computación.* Editado por Miguel Arias Estrada y Alexander Gelbuch. Colima: Sociedad Mexicana de Ciencias de la Computación: 219-226.

Girju, R. 2003. “Automatic Detection of Causal Relations for Question Answering”. *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*: 76-83.

Girju, R., Badulescu, A., Moldovan, D.2006: “Automatic Discovery of Part–Whole Relations”. *Computational Linguistics*, 32(1): 83-135.

Goddard, C., Wierzbicka, A. (eds.). 1994. *Semantic and Lexical Universals - Theory and Empirical Findings*. Amsterdam/Philadelphia: John Benjamins.

Gruber, T. 1993: “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”. In *Formal Ontology in Conceptual Analysis and Knowledge Representation*.

Guarino, N. 1998: “Formal Ontology and Information Systems”. *Proceedings of FOIS’98*, Trento, Italia. Amsterdam, IOS Press: 3-15.

Guzmán, R., Rosso, P., Montes, M. Villaseñor, L., Pinto, D. 2009: "Semi-supervised Word Sense Disambiguation Using the Web as Corpus". CICLing 2009: 256-265.

Harris, Z. 1970: *Distributional Structure*. In Papers in structural and Transformational Linguistics: 775-794.

Hearst, M. 1992: "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *Proceedings of COLING-92*, Nantes, France: 539-545.

Heid, U. 1999. "A Linguistic Bootstrapping Approach to the Extraction of Term Candidates from German Text". *Terminology* Vol. 5(2), John Benjamins: Amsterdam: 161-181.

Hirst, G. 2004: *Handbook on ontologies*. Editors S. Staab, R. Studer. Springer Verlag.

Hofweber, T. 2004: "Logic and Ontology". *Stanford Encyclopaedia of Philosophy.* Disponible en: http:\\plato.stanford.edu\entries\logic-ontology.

Jurafsky, D., Martin, J. 2009: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall.

Kant, I. 1800/1988: *Logik: Ein Handbuch zu Vorlesungen*. [Translated as Logic by R. S. Hartmann and W. Schwarz] (Dover Publications, New York).

Katz, J., Fodor, J. 1963: *The structure of a Semantic Theory*. Language 39: 170-210.

Krifka, M. 1998: *Lexical Relations*. LIN393S. Lexical Semantics.

Laurence, S., Margolis, E.1999: "Concepts and Cognitive Science". *Concepts: Core Readings*. Massachusetts Institute of Technology.

Lakoff, G. 1987: *Women, fire and dangerous things: What categories reveal about the mind*. The University of Chicago Press.

Leibniz, G. 1666: "Dissertatio de arte combinatoria", Leibnizens mathematische Schriften 5, Georg Olms, Hildesheim.

Lesk, M. 1986. “Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Crem Cone”. In *ACM SIGDOC Conference, ACM Press*: 24-26.

Levy-Bruhl, L. 1926. *How natives think*. [Fonctions mentales dans les societies inférieures, Paris, 1912]. Londres, Allen & Unwin.

Lin, D., Pantel, P. 2001: “Induction of Semantic Classes from Natural Language Text”. In: *KDD’01 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*: 317-322.

Litkowski, K. 2002: “Digraph Analysis of Dictionary Preposition Definitions”. *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia. Association for Computational Linguistics.

Lyons, J. 1977: *Semantics*. Cambridge University Press.

Lyons, J. 1968. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.

Llull, R. 1303: *Lògica Nova*. Vol. 4 of Nova Edició de les Obres de Ramon Llull, ed. by A. Bonner, Palma de Mallorca, 1998.

Malaisé, V. 2005: *Méthodologie linguistique et terminologique pour la structuration d’ontologies différentielles á partir de corpus textuels*. Paris, Université Paris 7–Denis Diderot.

Masterman, M. 1961. “Translation”, *Proceedings of the Aristotelian Society*: 170–216.

Mill, J.S. 1865. *A System of Logic*. (Longmans, London).

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. 1990. “WordNet: an on-line lexical database”. *Special issue of International Journal of Lexicography* 3(4): 235–44.

Miller, G. 1998: “Nouns in WordNet”. In *C. Fellbaum (ed.), WordNet: an electronic lexical database*. Cambridge, MA: MIT Press, 23-46.

Miller, J., Torii, M., Vijay, K. 2007: “Adaptation of POS tagging for Multiple Biomedical domains”. In: *BioNLP '07 Proceedings of the Workshop on BioNLP*

2007: Biological, Translational, and Clinical Language Processing: 179-180.

Minsky, M. 1975: "A Framework for Representing Knowledge". Reprinted in *The Psychology of Computer Vision*, P. Winston (Ed.), McGraw-Hill, 1975

Monachesi, P. 2007. "The LT4eL Project: Overview". [En línea]. Utrecht, Universidad de Utrecht. www.lt4el.eu/content/files/ws_prague/lt4el-prague.pdf

Muresan, S. y Klavans, J. 2002. "A Method for Automatically Building and Evaluating Dictionary Resources". In *Proceedings of the 3th International Conference on LanguageResources and Evaluation (LREC'02)*. Las Palmas, 29 a 31 de mayo: 231-234.

Murphy, G. 2002: *The big book of concepts*. Massachusetts Institute of Technology.

Murphy, L. 2003: *Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms*. Cambridge, University Press.

Navigli, R., Velardi, P. 2007: "GlossExtractor: A Web Application to Automatically Create a Domain Glossary". *Lecture Notes in Computer Science* 4733: 339-349.

Noy, N., McGuinness, D. 2001: "Ontology Development 101: A Guide to Creating your First Ontology". *Knowledge Systems Laboratory, Stanford University, CA*.

Ortega, R., Montes, M., Villaseñor, L. 2007: "Using Lexical Patterns for Extracting Hyponyms from the Web". In: *MICAI 2007. Advances in Artificial Intelligence. LNCS*, Vol. 4827, Springer, Berlin: 904-911.

Pajares, G., Santos, M. 2006: *Inteligencia Artificial e Ingeniería del Conocimiento*. Alfaomega grupo editor.

Pantel, P., Pennacchiotti, M. 2006: "Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations". *Proceedings of Conference on Computational Linguistics Association for Computational Linguistics*, Sydney, Australia: 113-120.

Pereira, F., Lee, L., Tishby, N. 1993: "Distributional Clustering of English Words". *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Ohio State University, Columbus, Ohio: 183-190.

Pinker, S. 1999: *Words and rules*. Weindefeld & Nicholson, London.

Pinto, D., Rosso, P., Jiménez, H. 2007: "UPV-SI: Word Sense Induction Using Self Term Expansion". In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007):* 430-433.

Poesio, M. 2005: "Domain Modelling and NLP: Formal Ontologies? Lexica? Or a Bit of Both?". *Applied Ontology Journal*. Volume 1 Issue 1, IOS Press Amsterdam, The Netherlands, The Netherlands.

Pugeault, F., Saint-Dizier, P., Monteil, M. 1994: "Knowledge Extraction from Texts: A Method for Extracting Predicate-Argument Structures from Texts". In *Proceedings of the 15th Conference on Computational Linguistics*, volume 2: 1039-1042.

Purandare, A., Pedersen, T. 2004: "Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces". In *Proceedings of the Conference on Computational Natural Language Learning*: 41-48.

Pustejovsky, J. 1991: "The Generative Lexicon". Computational Linguistics. Vol. 17, Issue 4. 409-441.

Quillian, M. 1968. "Semantic Memory". In M. Minsky (ed.), *Semantic Information Processing*, MIT Press: 227-270.

Rebeyrolle, J., Tanguy, L. 2000: "Repérage automatique de structures linguistiques en corpus: le cas des énoncés définitoires". *Cahiers de Grammaire* 25: 153-174.

Riloff, E., Shepherd, J.2004: "A corpus-based bootstrapping algorithm for Semi-automated semantic lexicon construction". *Journal of Natural Language Engineering,* 5(2): 147-156.

Riloff, E., Lorenzen, J. 1999. "Extraction-based text categorization: Generating domain-specific role relationships automatically". In Tomek Strzalkowski (Ed.), *Natural Language Information Retrieval.* Dordrecht, The Netherlands: Kluwer Academic Publishers: 167-196.

Rodríguez, C. 2004. “Metalinguistic Information Extraction for Terminology”. En *Proceedings of the 3rd International Workshop on Computational Terminology (CompuTerm2004)*. Génova, 29 de agosto: 15-22.

Rosch, E., Lloyd, B. 1978: *Cognition and categorization*. Hillsdale, Nueva York, Eribaum.

Rosch, E., Mervis, C. 1975: “Family resemblances. Studies in the internal structure of categories”. *Cognitive Psychology* 7: 573-605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., Boyes Braem, P.1976: “Basic objects in natural categories”. *Cognitive Psychology*, 8: 382-439.

Rosch, E. 1978: “Principles of categorization”. In Eleanor Rosh and Barbara B. Lloyd, editors, *Cognition and Categorization*, Lawrence Erlbaum Associates, Hillsdale, New Jersey. Chapter 2: 27-48.

Ryu, K., Choy, P. 2005: “An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning”. In: Buitelaar, P., Cimiano, P., & Magnini, B. (eds.*) Ontology Learning from Text: Methods, Evaluation and Applications*, IOS Press, Amsterdam: 15-28.

Saggion, H. 2004: “Identifying Definitions in Text Collections for Question Answering”. En *Proceedings 4th International Conference on Language Resources and Evaluation LREC2004*. Lisboa, 26 a 30 de mayo: 1927-1930.

Sánchez, A., Márquez, M. 2005: “Hacia un sistema de extracción de definiciones en textos jurídicos”. En *Actas de la 1er Jornada Venezolana de Investigación en Lingüística e Informática*. Venezuela, 14 de Octubre: 1-10.

Sager, J. C. 1990: *A Practical Course in Terminology Processing*. John Benjamins, Philadelphia Amsterdam.

Sarmento, L.; Maia, B.; Santos, D.; Pinto, A. y Cabral, L. 2006. “Corpógrafo V3. From Terminological Aid to Semi-automatic 244 Knowledge Engineering”. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. Génova, 22 a 28 de mayo: 1502-1505.

Saussure, F. 1968: *Curso de Lingüística General*, trad., prol. y notas de Amado Alonso, Losada, Buenos Aires.

Schank, R. 1975: *Conceptual Information Processing.* North-Holland Publishing Company, Amsterdam.

Schmid, H. 1994: "Probabilistic Part-of-Speech Tagging Using Decision Trees". *Proceedings of International Conference on New Methods in Language Processing*, September (1994), www.ims.uni-stuttgart.de~schmid.

Schütze, H. 1998: "Automatic Word Sense Discrimination". *Computational Linguistics,* 24(1): 97-123.

Segura, I., Martínez, J., Martínez, P. 2006: "Una Propuesta para el Etiquetado Automático de Roles Semánticos". *Procesamiento de lenguaje natural, No. 37:* 309-316.

Sierra, G., Alarcon, R., Aguilar, C., Bach, C. 2008: "Definitional verbal patterns for semantic relation extraction". *Terminology*, 14(1): 74-98.

Smith, E., Medin, D. 1981: *Categories and Concepts*. Harvard University Press. Cambridge, Massachusetts. London, England.

Smith, B. 2004: "Beyond Concepts: Ontology as Reality Representation". *Formal Ontology in Information Systems*. A.C. Varzi and I. Vieu (Eds.). IOS Press.

Smith, E.1988: *The psychology of human thought*. Cambridge University Press.

Snow, R., Jurafsky, D., Ng, A. 2005: "Learning Syntactic Patterns for Automatic Hypernym Discovery". L.K. Saul, Y. Weiss, L. Bottou (Eds.), *Advances in Neural Information Processing Systems*, vol. 17. 17MIT Press, Cambridge, MA.

Sowa, J. 2005: "Categorization in Cognitive Computer Science". In *Handbook of Categorization in Cognitive Science*. Edited by Henri Cohen and Claire Lefebvre. Elsevier Ltd.

Storrer, A., Wellinghoff, S. 2006. "Automated Detection and Annotation of Term Definitions in German Text Corpora". En *Proceedings of the 5th International Conference on LanguageResources and Evaluation (LREC'06)*. Génova: 2373-2376.

Sussna, M. 1993: "Word Sense Disambiguation for Free-test Indexing Using a Massive Semantic Network". In *2nd International Conference on Information and Knowledge Management:* 67-74.

Tanaka, J., Taylor, M. 1991: “Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder?”. *Cognitive Psychology*, 15: 121–149.

Velardi, P., Fabriani, P., Missikoff. 2001: “Using Text Processing Techniques to Automatically Enrich a Domain Ontology”. *Proceedings of the ACM International Conference on Formal Ontology in Information Systems*.

Vivaldi, J. 2001: *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*. Tesis doctoral. Universidad Politécnica de Catalunya, Departament de Llenguatges i Sistemes Informatics.

Whewell, W. 1858: *History of Scientific Ideas*. J.W. Parker & Son, London.

Wierzbicka, A. 1996: *Semantics: Primes and Universals*. Oxford University Press.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Wilks, Y., Slator, B., Guthrie, L. 1996: *Electric Words: dictionaries, computers and Meanings*. Cambridge, MA, MIT Press.

Winston, M., Chaffin, R., Herrman, D. 1987: “A Taxonomy of Part-Whole Relations”. *Cognitive Science* 11: 417-444.

Yarowsky, D. 1992: “Word-sense Disambiguation Using Statistical Models of Rogets Categories Trained on Large Corpora”. In *14th Conference on Computational Linguistics:* 454-460.

Zhang, F., Shi, S., Liu, J., Sun, S., Lin, Ch. 2011: “Nonlinear Evidence Fusion and Propagation for Hyponymy Relation Mining”. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon: 1159-1168.

Apéndices

Tabla 36. Etiquetas del etiquetador de partes de la oración TreeTagger para el español.

Etiqueta	Descripción
ACRNM	Acrónimo (ISO, CEI)
ADJ	Adjetivos (mayores, mayor
ADV	Adverbios (muy, demasiado, cómo)
ALFS	Letra singular del alfabeto
ART	Artículos (un, las, la, unas)
BACKSLASH	\
CARD	Cardinales
CC	Conjunción coordinativa (y, o)
CCAD	Conjunción coordinativa adversativa (pero)
CM	Coma (,)
COLON	(:)
CQUE	Que (como conjunción)
DASH	-
CSUBI	Conjunción subordinante que introduce una oración infinitiva
CSUBX	Conjunción subordinante subespecificada
DM	Pronombres demostrativos
FS	Puntuación de alto completo (.)
NC	Nombre común
NP	Nombre propio
ORD	Ordinales
PAL	Palabra formada por la contracción de “a” y “el”
PDEL	Palabra formada por la contracción de “de” y “el”
PE	Palabra extranjera
PPC	Pronombre personal clítico (le, les)
PREP	Preposición
REL	Pronombres relativos (cuyas, cuyo)
SE	Se (como partícula)
SEMICOLON	Punto y coma
SLASH	/
VSfin	Verbo ser
VSinf	Verbo ser conjugado
VLadj	Verbo en participio pasado
VLfin	Verbo conjugado
VLinf	Verbo en infinitivo
Vmfin	Verbo modal

Tabla 37. Hiperónimos candidatos y los CDs relevantes. Corpus Medicina.

Hiperónimo	**Frecuencia**	**CDs relevantes**	**Hiperónimo**	**Frecuencia**	**CDs relevantes**
Enfermedad	58	47	Fenómeno	4	3
Trastorno	49	48	Dolor	4	4
Examen	40	40	Estudio	4	3
Afección	29	29	Cambio	4	3
Procedimiento	22	22	Ruido	4	4
Infección	19	18	Acumulación	4	4
Proteína	17	16	Bloqueo	4	4
Cáncer	16	16	Presión	4	4
Tumor	16	15	Emergencia	4	0
Tratamiento	15	13	Mecanismo	4	0
Cirugía	14	12	Área	3	1
Método	12	11	Daño	3	3
Reservorio	12	0	País	3	0
Problema	12	11	Característica	3	1
Factor	11	0	Persona	3	2
Proceso	10	9	Condición	3	3
Inflamación	10	10	Sensación	3	3
Término	10	0	Vacuna	3	3
Complicación	10	10	Molestia	3	3
Medicamento	9	9	Antibiótico	3	3
Órgano	9	9	Unidad	3	0
Agente	9	9	Bacilo	3	2
Sustancia	8	8	Micosis	3	1
Glándula	8	8	Mineral	3	3
Dispositivo	8	8	Hallazgo	3	0
Estrechamiento	7	7	Medida	3	2
Elemento	7	2	Medición	3	3
Técnica	7	7	Hormona	3	3
Prueba	7	7	Fármaco	3	3
Síndrome	6	5	Categoría	3	3
Hinchazón	6	6	Molécula	3	3
Fuente	6	0	Capacidad	3	3
Clave	6	0	Lesión	3	3
Tubo	6	3	Instrumento	3	2
Defecto	5	5	Especie	3	0
Cavidad	5	5	Alteración	3	3
Herramienta	5	4	Urgencia	3	0
Opción	5	0	Arritmia	3	0
Radiación	4	4	Huésped	3	0
Partícula	5	5	Sistema	3	1
Presencia	5	0	Necrosis	3	3
Conducto	5	3			

Tabla 38. Hiperónimos extraídos con la expresión “Tipo de”.

Hiperónimo	**Frecuencia**
Cáncer	54
Medicamento	23
Tumor	17
Cirugía	15
Enfermedad	14
Virus	14
Célula	12
Tratamiento	12
Problema	11
Examen	10
Vacuna	10
Infección	9
Lesión	8
Calzado	7
Grasa	6
Dolor	5
Bacteria	5
Fármaco	5
Demencia	5
Angina	5
Ejercicio	4
Cálculo	4
Estructura	4
Prueba	4
Complicación	4
Proteína	4
Preparación	4
Prostatitis	4
Dieta	4
Necrosis	4
Inmunidad	4
Población	4

Tabla 39. Hiperónimo enfermedad y sus modificadores adjetivos y nominales. Resultados sin filtro. Corpus de medicina.

Hiperónimo: Enfermedad/nc, ocurrencia 1707 , Número total de relaciones 243
visceral/nc 2.0 Frec.Conj: 1 PMI: 0.281
autoinmun/adj 13.0 Frec.Conj: 7 PMI: 0.348
prodromic/adj 11.0 Frec.Conj: 1 PMI: 0.131
cromosomico/adj 28.0 Frec.Conj: 2 PMI: 0.117
septicemic/adj 8.0 Frec.Conj: 1 PMI: 0.159
alarmant/adj 6.0 Frec.Conj: 1 PMI: 0.184
afecto/adj 19.0 Frec.Conj: 7 PMI: 0.307
cerebral/adj 261.0 Frec.Conj: 1 PMI: -0.149
neoplasic/adj 3.0 Frec.Conj: 1 PMI: 0.246
causal/adj 76.0 Frec.Conj: 1 PMI: -0.04
cardiac/adj 1643.0 Frec.Conj: 153 PMI: 0.239
asintomatic/adj 75.0 Frec.Conj: 1 PMI: -0.039
neonatal/nc 2.0 Frec.Conj: 1 PMI: 0.281
mortal/adj 37.0 Frec.Conj: 7 PMI: 0.236
neurologic/adj 95.0 Frec.Conj: 6 PMI: 0.117
ocular/adj 171.0 Frec.Conj: 10 PMI: 0.115
cerebrovascular/adj 142.0 Frec.Conj: 9 PMI: 0.123
drepanocitic/nc 14.0 Frec.Conj: 4 PMI: 0.264
virulento/adj 13.0 Frec.Conj: 1 PMI: 0.116
grave/adj 382.0 Frec.Conj: 41 PMI: 0.217
microbia./nc 2.0 Frec.Conj: 2 PMI: 0.365
malig./nc 3.0 Frec.Conj: 1 PMI: 0.246
quistic/adj 40.0 Frec.Conj: 5 PMI: 0.185
multifactorial/adj 8.0 Frec.Conj: 3 PMI: 0.284
estafilococi/nc 4.0 Frec.Conj: 4 PMI: 0.39
pulmonar/nc 17.0 Frec.Conj: 2 PMI: 0.164
postrombotica/adj 2.0 Frec.Conj: 1 PMI: 0.281
prematuro/adj 79.0 Frec.Conj: 2 PMI: 0.019
inmunosupresor/adj 15.0 Frec.Conj: 2 PMI: 0.175
diarreic/adj 26.0 Frec.Conj: 19 PMI: 0.426
cardiac/nc 34.0 Frec.Conj: 2 PMI: 0.098
autoinmunitari/adj 30.0 Frec.Conj: 14 PMI: 0.359
tasa/nc 68.0 Frec.Conj: 1 PMI: -0.03
aterosclerotica/adj 4.0 Frec.Conj: 3 PMI: 0.351
hemorragic/adj 103.0 Frec.Conj: 4 PMI: 0.064
sintomatic/adj 49.0 Frec.Conj: 6 PMI: 0.187
glomerular/adj 23.0 Frec.Conj: 3 PMI: 0.18
frecuente/adj 79.0 Frec.Conj: 1 PMI: -0.043
linfoproliferativ/adj 5.0 Frec.Conj: 1 PMI: 0.201
tubercul/nc 227.0 Frec.Conj: 8 PMI: 0.058
arterial/adj 677.0 Frec.Conj: 7 PMI: -0.074
invasor/adj 36.0 Frec.Conj: 22 PMI: 0.412
transmisible/adj 62.0 Frec.Conj: 59 PMI: 0.529
digestiv/adj 62.0 Frec.Conj: 8 PMI: 0.198
agente/nc 484.0 Frec.Conj: 1 PMI: -0.203
bacteriemic/adj 4.0 Frec.Conj: 1 PMI: 0.22
febril/adj 52.0 Frec.Conj: 17 PMI: 0.325
similar/adj 227.0 Frec.Conj: 7 PMI: 0.043
tuberculoide/nc 7.0 Frec.Conj: 1 PMI: 0.171
resistente/adj 62.0 Frec.Conj: 1 PMI: -0.022

tromboembolic/adj 2.0 Frec.Conj: 1 PMI: 0.281
sever/adj 87.0 Frec.Conj: 1 PMI: -0.052
endocrin/adj 33.0 Frec.Conj: 2 PMI: 0.101
profesional/adj 47.0 Frec.Conj: 1 PMI: 0.003
focal/adj 7.0 Frec.Conj: 1 PMI: 0.171
oseo/adj 281.0 Frec.Conj: 4 PMI: -0.037
inflamatori/adj 115.0 Frec.Conj: 38 PMI: 0.361
pulmonar/adj 956.0 Frec.Conj: 121 PMI: 0.278
erradicable/adj 1.0 Frec.Conj: 1 PMI: 0.343
clasico/adj 48.0 Frec.Conj: 1 PMI: 0.001
hidatidic/adj 2.0 Frec.Conj: 1 PMI: 0.281
estomacal/adj 77.0 Frec.Conj: 1 PMI: -0.041
prolongado/adj 20.0 Frec.Conj: 1 PMI: 0.078
isquemic/adj 93.0 Frec.Conj: 2 PMI: 0.004
periodontal/adj 14.0 Frec.Conj: 7 PMI: 0.34
repentin/adj 52.0 Frec.Conj: 1 PMI: -0.006
epidemic/adj 59.0 Frec.Conj: 2 PMI: 0.047
neuromuscular/adj 6.0 Frec.Conj: 2 PMI: 0.262
determinado/adj 26.0 Frec.Conj: 1 PMI: 0.055
sujeto/adj 4.0 Frec.Conj: 2 PMI: 0.3
conexo/adj 9.0 Frec.Conj: 3 PMI: 0.272
renal/nc 633.0 Frec.Conj: 44 PMI: 0.161
agravant/adj 4.0 Frec.Conj: 1 PMI: 0.22
enteric/adj 48.0 Frec.Conj: 3 PMI: 0.108
fulminant/adj 19.0 Frec.Conj: 4 PMI: 0.234
alergic/adj 90.0 Frec.Conj: 3 PMI: 0.047
natural/adj 177.0 Frec.Conj: 9 PMI: 0.099
metabolic/adj 56.0 Frec.Conj: 2 PMI: 0.052
tratabl/adj 1.0 Frec.Conj: 1 PMI: 0.343
universal/adj 32.0 Frec.Conj: 1 PMI: 0.037
silencios/adj 13.0 Frec.Conj: 1 PMI: 0.116
viral/adj 40.0 Frec.Conj: 5 PMI: 0.185
invasiv/adj 43.0 Frec.Conj: 7 PMI: 0.22
gastric/adj 157.0 Frec.Conj: 1 PMI: -0.104
mental/adj 175.0 Frec.Conj: 28 PMI: 0.256
antimembrana/adj 8.0 Frec.Conj: 3 PMI: 0.284
dermatologic/adj 7.0 Frec.Conj: 1 PMI: 0.171
aortic/adj 194.0 Frec.Conj: 1 PMI: -0.123
semejante/adj 20.0 Frec.Conj: 2 PMI: 0.148
fisico/adj 486.0 Frec.Conj: 1 PMI: -0.204
dominante/adj 12.0 Frec.Conj: 1 PMI: 0.123
endocrinometabolic/adj 3.0 Frec.Conj: 1 PMI: 0.246
tuberculos/adj 3.0 Frec.Conj: 1 PMI: 0.246
infiltrativ/nc 3.0 Frec.Conj: 1 PMI: 0.246
estructural/adj 60.0 Frec.Conj: 2 PMI: 0.045
hipertensiv/nc 32.0 Frec.Conj: 6 PMI: 0.231
bulos/adj 7.0 Frec.Conj: 1 PMI: 0.171
respiratori/adj 11.0 Frec.Conj: 1 PMI: 0.131
zoonotic/adj 8.0 Frec.Conj: 1 PMI: 0.159
cardiovascular/adj 216.0 Frec.Conj: 73 PMI: 0.397
pleural/adj 80.0 Frec.Conj: 1 PMI: -0.044
inmunitari/adj 252.0 Frec.Conj: 1 PMI: -0.146
temprano/adj 182.0 Frec.Conj: 2 PMI: -0.059
causante/nc 30.0 Frec.Conj: 2 PMI: 0.11
renal/adj 2.0 Frec.Conj: 1 PMI: 0.281

renales/nc 136.0 Frec.Conj: 4 PMI: 0.036
extrahepatic/adj 1.0 Frec.Conj: 1 PMI: 0.343
llamada/nc 125.0 Frec.Conj: 1 PMI: -0.084
atribuibl/adj 9.0 Frec.Conj: 2 PMI: 0.224
fantasmo/adj 2.0 Frec.Conj: 1 PMI: 0.281
ateroembolic/adj 11.0 Frec.Conj: 1 PMI: 0.131
trasmisibl/adj 10.0 Frec.Conj: 10 PMI: 0.43
hepatic/adj 3.0 Frec.Conj: 2 PMI: 0.327
granulomatos/nc 13.0 Frec.Conj: 2 PMI: 0.189
epidemiologicamente/adj 4.0 Frec.Conj: 1 PMI: 0.22
mitro-aortic/nc 1.0 Frec.Conj: 1 PMI: 0.343
poliquistic/adj 15.0 Frec.Conj: 2 PMI: 0.175
coronari/nc 272.0 Frec.Conj: 14 PMI: 0.105
nosocomial/adj 27.0 Frec.Conj: 3 PMI: 0.165
recurrente/adj 40.0 Frec.Conj: 1 PMI: 0.017
intestinal/adj 159.0 Frec.Conj: 13 PMI: 0.157
ocupacional/adj 42.0 Frec.Conj: 2 PMI: 0.079
ligero/adj 31.0 Frec.Conj: 1 PMI: 0.039
comun/adj 291.0 Frec.Conj: 7 PMI: 0.016
preexistent/adj 13.0 Frec.Conj: 2 PMI: 0.189
fungic/adj 4.0 Frec.Conj: 1 PMI: 0.22
infantil/adj 115.0 Frec.Conj: 3 PMI: 0.023
hepatic/nc 109.0 Frec.Conj: 19 PMI: 0.255
reactiv/adj 13.0 Frec.Conj: 1 PMI: 0.116
resecabl/adj 1.0 Frec.Conj: 1 PMI: 0.343
clinic/adj 1080.0 Frec.Conj: 49 PMI: 0.106
broncopulmonar/adj 9.0 Frec.Conj: 1 PMI: 0.149
valvular/adj 85.0 Frec.Conj: 1 PMI: -0.05
cronic/adj 533.0 Frec.Conj: 62 PMI: 0.24
venos/adj 113.0 Frec.Conj: 4 PMI: 0.054
varicos/adj 14.0 Frec.Conj: 2 PMI: 0.182
neuroparalitic/adj 1.0 Frec.Conj: 1 PMI: 0.343
esporadico/adj 73.0 Frec.Conj: 12 PMI: 0.235
diverticular/adj 6.0 Frec.Conj: 6 PMI: 0.407
matern/adj 114.0 Frec.Conj: 1 PMI: -0.076
actual/adj 145.0 Frec.Conj: 4 PMI: 0.029
previo/adj 196.0 Frec.Conj: 2 PMI: -0.066
polimorf/nc 3.0 Frec.Conj: 1 PMI: 0.246
degenerativ/nc 13.0 Frec.Conj: 1 PMI: 0.116
extracraneal/adj 1.0 Frec.Conj: 1 PMI: 0.343
tardi/adj 52.0 Frec.Conj: 1 PMI: -0.006
anergic/nc 1.0 Frec.Conj: 1 PMI: 0.343
mundial/adj 281.0 Frec.Conj: 1 PMI: -0.155
seudogripal/adj 3.0 Frec.Conj: 1 PMI: 0.246
parasitari/adj 19.0 Frec.Conj: 5 PMI: 0.262
gestacional/nc 4.0 Frec.Conj: 1 PMI: 0.22
propuest/nc 4.0 Frec.Conj: 1 PMI: 0.22
miocardic/adj 71.0 Frec.Conj: 2 PMI: 0.029
tropical/adj 61.0 Frec.Conj: 3 PMI: 0.085
objeto/nc 49.0 Frec.Conj: 11 PMI: 0.267
catastrofic/adj 2.0 Frec.Conj: 1 PMI: 0.281
viric/nc 229.0 Frec.Conj: 17 PMI: 0.151
intercurrent/adj 10.0 Frec.Conj: 4 PMI: 0.298
raro/adj 70.0 Frec.Conj: 10 PMI: 0.214
bacterian/adj 156.0 Frec.Conj: 25 PMI: 0.253

oclusiv/adj 3.0 Frec.Conj: 1 PMI: 0.246
estacional/adj 9.0 Frec.Conj: 2 PMI: 0.224
neuromuscular/nc 1.0 Frec.Conj: 1 PMI: 0.343
especifico/adj 595.0 Frec.Conj: 12 PMI: -0.003
incapacitant/adj 2.0 Frec.Conj: 1 PMI: 0.281
exclusivo/adj 5.0 Frec.Conj: 1 PMI: 0.201
coexistent/adj 3.0 Frec.Conj: 1 PMI: 0.246
respiratori/nc 586.0 Frec.Conj: 7 PMI: -0.058
particular/adj 83.0 Frec.Conj: 3 PMI: 0.055
neumococic/adj 32.0 Frec.Conj: 11 PMI: 0.315
autosomico/adj 21.0 Frec.Conj: 4 PMI: 0.224
inmunoprevenibl/adj 3.0 Frec.Conj: 3 PMI: 0.38
sistemic/adj 100.0 Frec.Conj: 15 PMI: 0.23
aparente/adj 19.0 Frec.Conj: 1 PMI: 0.083
abdominal/nc 192.0 Frec.Conj: 1 PMI: -0.122
agudo/adj 559.0 Frec.Conj: 22 PMI: 0.078
subyacente/adj 67.0 Frec.Conj: 9 PMI: 0.205
infectocontagios/adj 2.0 Frec.Conj: 2 PMI: 0.365
diafragmatic/nc 10.0 Frec.Conj: 1 PMI: 0.139
inmunodepresor/adj 5.0 Frec.Conj: 1 PMI: 0.201
venere/adj 15.0 Frec.Conj: 3 PMI: 0.222
metastasic/adj 5.0 Frec.Conj: 2 PMI: 0.279
familiar/adj 195.0 Frec.Conj: 1 PMI: -0.123
prevenibl/adj 28.0 Frec.Conj: 26 PMI: 0.472
vacuna/nc 796.0 Frec.Conj: 3 PMI: -0.166
significativo/adj 103.0 Frec.Conj: 1 PMI: -0.067
activo/adj 194.0 Frec.Conj: 22 PMI: 0.207
leve/adj 128.0 Frec.Conj: 11 PMI: 0.16
importante/adj 284.0 Frec.Conj: 5 PMI: -0.016
concomitant/adj 14.0 Frec.Conj: 2 PMI: 0.182
benigno/adj 81.0 Frec.Conj: 12 PMI: 0.223
indicador/adj 6.0 Frec.Conj: 5 PMI: 0.381
organic/adj 33.0 Frec.Conj: 1 PMI: 0.034
psiquiatric/adj 3.0 Frec.Conj: 1 PMI: 0.246
periodonta./nc 1.0 Frec.Conj: 1 PMI: 0.343
exotic/adj 3.0 Frec.Conj: 1 PMI: 0.246
meningococic/adj 21.0 Frec.Conj: 8 PMI: 0.315
colorrectal/adj 61.0 Frec.Conj: 1 PMI: -0.02
monogenic/nc 4.0 Frec.Conj: 2 PMI: 0.3
contagios/adj 7.0 Frec.Conj: 2 PMI: 0.247
hereditario/adj 33.0 Frec.Conj: 6 PMI: 0.228
infeccios/adj 515.0 Frec.Conj: 132 PMI: 0.391
genic/nc 10.0 Frec.Conj: 1 PMI: 0.139
nuevo/adj 172.0 Frec.Conj: 8 PMI: 0.088
tuberculos/nc 11.0 Frec.Conj: 2 PMI: 0.205
intrinseco/adj 22.0 Frec.Conj: 1 PMI: 0.07
cutane/adj 241.0 Frec.Conj: 4 PMI: -0.022
predisponent/adj 11.0 Frec.Conj: 1 PMI: 0.131
vascular/adj 200.0 Frec.Conj: 19 PMI: 0.182
terminal/adj 104.0 Frec.Conj: 1 PMI: -0.068
ulceros/adj 16.0 Frec.Conj: 2 PMI: 0.169
debilitant/adj 10.0 Frec.Conj: 3 PMI: 0.262
paludic/adj 18.0 Frec.Conj: 1 PMI: 0.087
extracardiac/adj 4.0 Frec.Conj: 1 PMI: 0.22
general/adj 470.0 Frec.Conj: 2 PMI: -0.149

primario/adj 306.0 Frec.Conj: 5 PMI: -0.024
pericardic/adj 17.0 Frec.Conj: 2 PMI: 0.164
celiac/nc 5.0 Frec.Conj: 4 PMI: 0.368
propio/adj 49.0 Frec.Conj: 2 PMI: 0.064
oportunist/adj 12.0 Frec.Conj: 1 PMI: 0.123
mitral/nc 199.0 Frec.Conj: 1 PMI: -0.125
supurativ/adj 3.0 Frec.Conj: 2 PMI: 0.327
human/adj 681.0 Frec.Conj: 5 PMI: -0.106
diferente/adj 148.0 Frec.Conj: 2 PMI: -0.04
genetic/adj 250.0 Frec.Conj: 10 PMI: 0.073
congenit/adj 142.0 Frec.Conj: 1 PMI: -0.095
progresiv/adj 59.0 Frec.Conj: 3 PMI: 0.088
costoso/adj 10.0 Frec.Conj: 1 PMI: 0.139
somatico/adj 12.0 Frec.Conj: 1 PMI: 0.123
reciente/adj 323.0 Frec.Conj: 4 PMI: -0.051
estructura/nc 79.0 Frec.Conj: 1 PMI: -0.043
endemic/adj 162.0 Frec.Conj: 9 PMI: 0.108
atopic/adj 5.0 Frec.Conj: 1 PMI: 0.201
distinto/adj 79.0 Frec.Conj: 3 PMI: 0.06
obstructiv/adj 26.0 Frec.Conj: 1 PMI: 0.055
inmunologic/adj 54.0 Frec.Conj: 2 PMI: 0.055
vectorial/adj 5.0 Frec.Conj: 1 PMI: 0.201
emergente/adj 70.0 Frec.Conj: 39 PMI: 0.43
miliar/nc 12.0 Frec.Conj: 1 PMI: 0.123
malign/adj 69.0 Frec.Conj: 6 PMI: 0.151
serio/adj 19.0 Frec.Conj: 1 PMI: 0.083
prototipo/nc 5.0 Frec.Conj: 1 PMI: 0.201
regional/adj 52.0 Frec.Conj: 1 PMI: -0.006