Recibido: 15 de marzo de 2016
Aceptado: 11 de mayo de 2016

MARCO ONTOLÓGICO PARA LA ESTRUCTURACIÓN SEMÁNTICA Y LA RECUPERACIÓN DE RECURSOS BIBLIOGRÁFICOS EMPLEANDO PROCESAMIENTO DEL LENGUAJE NATURAL

ONTOLOGICAL FRAMEWORK FOR SEMANTIC MARKUP AND BIBLIOGRAPHIC INFORMATION RETRIEVAL USING NATURAL LANGUAGE PROCESSING

PhD. Torcoroma Velásquez Pérez, MSc. (c) Andrés Mauricio Puentes
MSc. Edwin E. Espinel Blanco

Universidad Francisco de Paula Santander Ocaña, Facultad de Ingenierías.
Sede Algodonal Ocaña, Norte de Santander, Colombia.
Tel.: (+577) 569 0088, Fax: (+577) 569 0088, Ext. 182.
E-mail: {tvelasquezp, eeespinelb}@ufpso.edu.co, ampuentesv@misena.edu.co.

Resumen: El proyecto tiene como propósito crear un modelo ontológico que describa y relacione los elementos requeridos para el procesamiento del lenguaje natural en el dominio de las búsquedas bibliográficas semánticas. Esta propuesta será abordada como una investigación del tipo descriptiva bajo un enfoque mixto dado que se pretende describir de modo sistemático las características de un modelo que describe una problemática muy común que puede ser abordada desde una perspectiva tecnológica.

Palabras clave: Recuperación de la información, web semántica, procesamiento de lenguaje natural.

Abstract: The project aims to create a semantic model for describe and link elements for natural language processing at bibliographic semantic search. This proposal will be developed as a descriptive research on a mixed approach, describing systematically the features of a model that describe a common problematic from a technologic perspective.

Keywords: Information retrieval, semantic web, natural language processing.

1. INTRODUCCIÓN

Las ontologías constituyen un elemento fundamental en la arquitectura de la Web Semántica, y por ello, han cobrado gran protagonismo al interior de la comunidad mundial de desarrollo web que se preocupa por realizar aportes en aspectos relacionados con mejorar la interacción entre personas y computadores. Para que pueda llevarse a cabo la web semántica y eliminar de esta manera los problemas de interacción con la Web actual se necesita que el conocimiento de la Web este representado de forma que sea legible por los computadores, esté consensuado y sea reutilizable.

El presente trabajo pretende ser un aporte en el dominio de la información bibliográfica, debido a que la necesidad que tienen las personas de acceder a las fuentes de información como libros, artículos, revistas y demás en una biblioteca, es muy grande; esta necesidad no se ve del todo cubierta debido a

las falencias que presentan los sistemas de información existentes en la actualidad, que realizan una catalogación muy pobre de los recursos, lo que repercute en serias dificultades para el usuario que quiere acceder a estas fuentes de información, dificultades que tienen que ver con un costoso filtrado manual que se debe hacer de los resultados imprecisos obtenidos después de cualquier consulta.

2. DESARROLLO

2.1 Metodología

La población considerada en el estudio se conforma inicialmente de un grupo de personas que relatarán su experiencia antes y después de introducir mejoras en lo relacionado el tratamiento semántico del PLN. Esta población está conformada por estudiantes de ingeniería de sistemas, trabajadores en desarrollo de software, y una pequeña cantidad de usuarios no expertos que interactúan frecuentemente con algún buscador de internet. Para definir las muestras se aplicará una formula estadística que permitirá una mayor precisión relacionada con el análisis de la información.

Para el desarrollo del proyecto se define la identificación de los procesos y los actores relacionados con la publicación y recuperación de recursos en un Sistema de Información Bibliográfico, posteriormente se debe representar en una conceptualización en idioma español las estructuras gramaticales y sintácticas del dominio bibliográfico, se continua con el diseño de los algoritmos requeridos por un sistema de búsquedas para procesar entradas en lenguaje natural y la implementación en una herramienta tecnológica con la solución algorítmica planteada que permita acceder al conocimiento representado.

2.2 Conceptualización

Para el desarrollo de un modelo que permita la estimación del significado de las palabras en un contexto determinado es importante apoyarse en un conjunto de teorías y de algoritmos que fundamentan este aspecto del PLN, a continuación se describirán aspectos fundacionales de estas disciplinas

La terminología y el procesamiento del lenguaje natural utilizan conocimiento de un dominio, desde la generación de textos, a la localización de ontologías, pasando por la recuperación de información, la traducción automática o la traducción asistida (TAO), o la anotación lingüística basada en ontologías.

A través del mapa conceptual (figura 1) se representa el estado del arte abordado para este trabajo.

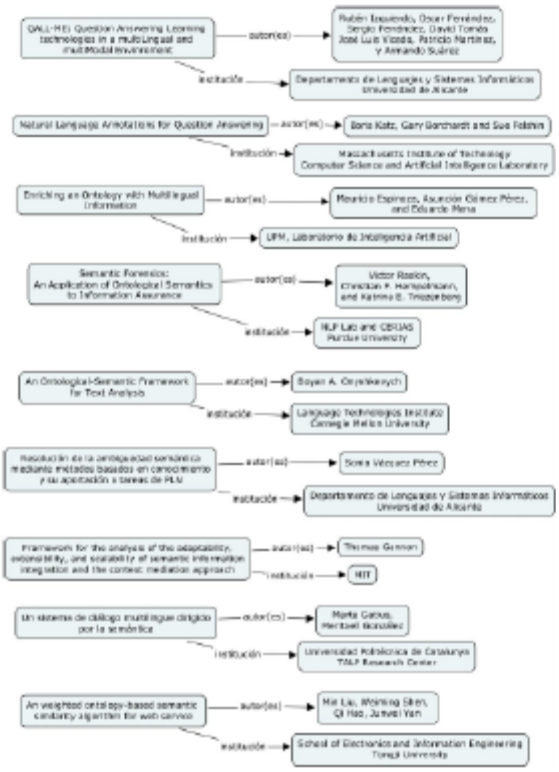

Fig. 1. Mapa Conceptual Estado del Arte

Es importante contar con conocimientos previos de Lingüística, Ciencia de la Computación, Lógica Matemática, Sicología Cognitiva, Filosofía del Lenguaje, Teoría de la Optimización, Estadística Bayesiana, Algebra lineal; lo cual convierte a este campo de investigación en multidisciplinar, y nos obliga a no tomar una visión sesgada al estudiar esta disciplina.

Existe una variedad de sistemas que podrían clasificarse como sistemas de PLN, sin embargo, el interés aquí se centra en los sistemas que tienen como entrada fragmentos del lenguaje escrito, es decir, texto en algún formato legible por un computador (figura 2). Existen otros sistemas que estudian lo concerniente al lenguaje hablado, pero estos incluyen una serie de características muy particulares relacionadas con la fonética y con otros aspectos que se salen del dominio de esta propuesta.

Fig. 2. Elementos del Lenguaje Natural

Los sistemas de información se pueden clasificar básicamente en una de tres categorías: Sistemas transaccionales, sistemas estratégicos y sistemas de apoyo a la toma de decisiones. Los sistemas basados en conocimiento han constituido una forma efectiva de implementar los mandatos del paradigma de los sistemas de apoyo a la toma de decisiones, debido a que las organizaciones gestionan más conocimiento que mera información.

La teoría y fundamentación de los SBC proviene de la Inteligencia Artificial, planteando una separación entre el conocimiento necesario para lograr los objetivos o metas y los algoritmos necesarios para la interpretación de este conocimiento. La complejidad de los sistemas presentes en el mundo nos ha guiado hacia nuevas formas de hacer ciencia, nuevas formas de buscar explicación y solución a las diversas situaciones caóticas del mundo actual, y la inteligencia artificial juega un papel fundamental en este contexto. Los sistemas basados en conocimiento han de contar con mecanismos que permitan no solamente representar conocimiento de algún dominio particular, sino también, proveer los medios y la tecnología para la recuperación efectiva de dicho conocimiento.

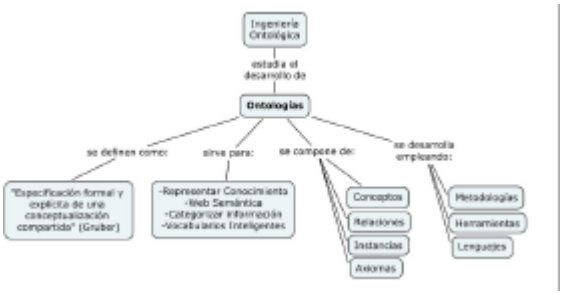

Fig. 3. Descripción Ingeniería Ontológica.

Dos pilares básicos de la IA son los principios que se deben seguir para la construcción de SBC: la hipótesis de representación de conocimiento que determina que el conocimiento se puede representar a través de unas estructuras de datos vistas como sentencias o proposiciones y la hipótesis de reflexión que determina la capacidad de un sistema inteligente de razonar sobre sí mismo gracias a un intérprete que manipula formalmente representaciones de su estructura y sus propias operaciones.

Las ontologías han surgido entonces como ese mecanismo efectivo para representar conocimiento de manera formal, consensuada, legible por computadores como soporte para la web semántica, cuyo éxito, según sus impulsores, se materializará por la disposición a compartir ontologías que muestren comunidades y grupos en la web.

Los sistemas QA (*Question Answering*) (figura 4) son un tipo específico de sistemas de recuperación de información que funcionan a través de preguntas y respuestas. Constituyen un avance importante hacia el desarrollo de aplicaciones de fácil interacción con el usuario porque presentan como salida una respuesta concreta a una pregunta realizada en lenguaje natural, en vez de proporcionar un listado de resultados que el usuario final debe filtrar manualmente, muchas veces sin éxito, representando un alto costo en tiempo y en recursos.

Los sistemas de QA requieren de muchos elementos en su arquitectura funcional para poder cumplir con su objetivo. En este trabajo se exploró la teoría que sustenta el funcionamiento de este tipo de sistemas, entendiéndolos como una aplicación importante de la IA donde confluyen diversos elementos que pertenecen al Procesamiento del Lenguaje Natural y a la Representación de Conocimiento. Esta última área de la Inteligencia Artificial juega un papel fundamental en la arquitectura de estos sistemas debido a la posibilidad que brinda de facilitar los diversos tipos de análisis que hay que realizar sobre el texto en lenguaje natural (análisis sintáctico, semántico, pragmático y de contexto) apoyándose en bases de conocimiento; es aquí donde toman importancia las ontologías como esas representaciones que aportan formalidad y especificidad a la representación del conocimiento de un dominio compartido.

Fig. 4. Representación de Sistemas QA.

2.3 Formalización

El dominio escogido para hacer la primera conceptualización de prueba fue el de los trabajos de fin de carrera o trabajos de grado; para esto se hizo un trabajo exhaustivo de consulta y abstracción de la información relevante tomando como piloto la información contenida en el Sistema de Información Bibliográfico de los estudiantes de la facultad de ingeniería de la Universidad Francisco de Paula Santander Ocaña; para esta labor se contó con la colaboración del personal que labora en la biblioteca y de los estudiantes de ingeniería de sistemas Natividad Galviz y Cindy Guerrero quienes apoyaron las labores de recolección de datos y sistematización de los resultados.

El resultado de esta primera fase fue la representación del conocimiento del dominio de los trabajos de grado en una base de conocimiento, implementada como una Ontología de nivel intermedio, desarrollada utilizando el editor Protégé y siguiendo un proceso de ingeniería muy detallado descrito en la metodología de desarrollo de ontologías llamada *Methontology*. Para el contexto de la ontología, con estos datos se logró estructurar y organizar un listado de los conceptos más importantes, de esta manera se determinan ocho clases principales que son: DirectorTrabajoDeGrado, EstudiantesDePregrado, Facultad, JuradoDeTrabajoDeGrado, LineaDeInvestigación, ModalidadDeTrabajoDeGrado, PlanDeEstudios, TrabajoDeGrado.

Tabla 1: Listado de Clases

DirectorTrabajoDeGrado
DirectorDeTrabajoDeGradoCursoDeProfundización
DirectorDeTrabajoDeGradoPasantia
DirectorDeTrabajoDeGradoTesis

Se realizó un estudio sobre las características generales de los trabajos de grado para determinar los rasgos distintivos de cada uno de ellos y posteriormente se elaboraron una serie de mapas conceptuales plasmando las características más sobresalientes de cada trabajo de grado. Se presentan las características más importantes que se pueden observar en un trabajo de grado (tabla 2).

Tabla 2: Características de un trabajo de grado

Numero de Registro	Signatura Topográfica	Código de Barras	EspecialidadDirector	Lineа de investigación
Título	Director	Modalidad	Año	Jurados
Autor	CódigoAutor	Plan de estudios	Resumen	Tipo Jurado

La web semántica se basa principalmente en mecanismos que permiten representar el conocimiento de un modo estandarizado, haciendo posible su tratamiento automático y la existencia por lo tanto, de numerosas nuevas aplicaciones que puedan beneficiarse de estas representaciones. La representación del conocimiento como tal, es una materia en la que se lleva trabajando desde hace varias décadas, desde mucho antes de que surgiera la web semántica. Surgió en el ámbito de la Inteligencia Artificial al tratar de crear representaciones de conocimiento que pudieran ser utilizadas por mecanismos que simulasen el razonamiento humano. En la ontología de trabajos de grado se usaron reglas como técnica para representar el conocimiento extraído, previamente organizado y clasificado de acuerdo a las características de los mismos. A continuación se muestra una regla (tabla 3).

Tabla 3: Ejemplo de representación de una regla

1.Lenguaje Natural	Un trabajo de grado (tesis), es realizado por un estudiante de pregrado en tesis y tiene una modalidad de trabajo de grado en tesis y pertenece a una facultad y pertenece a un plan de estudios entonces tiene una linea de investigación.
2. Lógica de Predicados	E es un estudiante TG es un trabajo de grado RealizoTrabajoDeGradoTesis(E, OntologiaPecesDelCatatumbo) ^ ^ TieneModalidadTrabajoDeGradoTesis(TG, Tesis) ^ PerteneceAUnaFacultad(TG, FacultadDeIngenieria) ^ PerteneceAUnPlanDeEstudio(TG, PlanDeEstudiosDeIngenieriaDeSistemas) → TieneUnaLineaDeInvestigacion(TG, LineaDeInvestigacionDeSistemasInteligentes)
3. SWRL	RealizoTrabajoDeGradoTesis(?x, OntologiaPecesDelCatatumbo) ^ EsRealizadoPorEstudianteDePregradoEnTesis(?x, LuisErnestoLopez) ^ TieneModalidadTrabajoDeGradoTesis(?x, Tesis) ^ PerteneceAUnaFacultad(?x, FacultadDeIngenieria) ^ PerteneceAUnPlanDeEstudio(?x, PlanDeEstudiosDeIngenieriaDeSistemas) → TieneUnaLineaDeInvestigacion(?x, LineaDeInvestigacionDeSistemasInteligentes)

En la formalización se transforma el modelo conceptual en un modelo formal o semi computable más cercano al lenguaje de implementación. Para lograrlo la tarea principal es seleccionar un sistema de representación del conocimiento como pueden ser: marcos o *Frames*, lógica descriptiva, modelamiento con UML, representaciones relacionales. Esta representación está inmersa en las reglas explicadas en la sección anterior.

2.4 Integración

El objetivo de la fase de integración es construir una Ontología reutilizando definiciones y conocimientos presentes en otras Ontologías, que serán integradas en la nueva Ontología en desarrollo. Esta integración debe realizarse durante todo el ciclo de vida, teniendo presente que es más significativa en la etapa de especificación y conceptualización que en la etapa de implementación. Ésta es una forma de aprovechar implementaciones ya realizadas, pero es muy

importante asegurar que las Ontologías a integrar cumplen ciertos requisitos de calidad. En este caso particular, no fue posible integrar otras ontologías debido la imposibilidad de encontrar implementaciones relacionadas con el área de estudio.

2.5 Implementación

El objetivo de esta fase es escribir la ontología en un lenguaje formal que sea computable. Se escoge OWL por ser uno de los lenguajes más importantes para la construcción de ontologías; su entorno de desarrollo se puede soportar en el editor de Ontologías Protégé 3.4. (Figura 5).

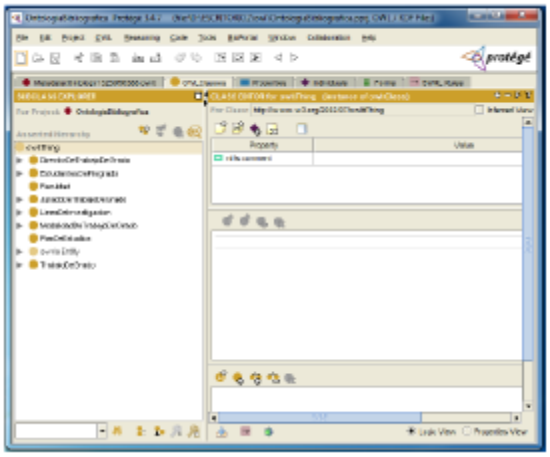

Fig. 5. Explorador de Clases y Subclases, Mostrando los Conceptos Principales

Las clases definidas en la clasificación taxonómica están agrupadas en DirectorTrabajoDeGrado, EstudiantesDePregrado, Facultad, JuradoDeTrabajoDeGrado, LineaDeInvestigacion, ModalidadDeTrabajoDeGrado, PlanDeEstudios y TrabajoDeGrado.

Para cada relación de define un rango y un dominio determinados. El rango indica los valores que puede tomar la relación y el dominio son las clases a las que se asigna dicha relación. En la figura se observa que la relación EsDirectorDeTrabajoDeGradoProfundizaciontiene como dominio DirectorDeTrabajoDeGradoCursoDeProfundizacion y como rango TrabajoDeGradoCursoDeProfundizacion.

Cada subclase de las clases principales está compuesta por instancias, estas últimas representan los datos actuales en la base de conocimiento de la ontología creada. Por ejemplo, la subclase EstudiantesDePregradoEnPasantia tiene 29 instancias las cuales se pueden observar en la parte derecha de la Figura (figura 6)

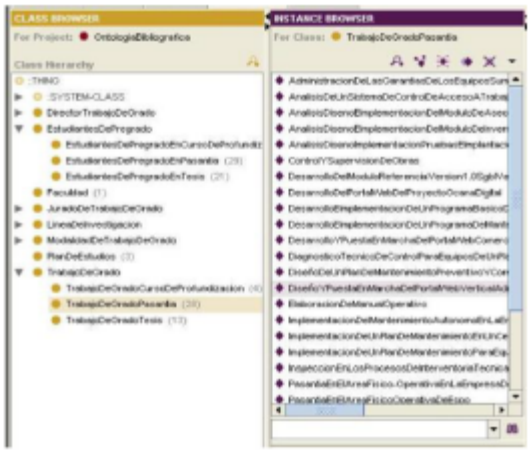

Fig. 6. Vista de Instancias TrabajoDeGradoPasantia

En la Figura se puede observar las instancias correspondientes a la subclase TrabajoDeGradoPasantia en este ejemplo la subclase está compuesta por 28 instancias. (fig. 7).

Fig. 7.Representación Gráfica de la Ontología de los Trabajos de Grado Utilizando OWL Viz Mostrando los Conceptos Principales y Subclases

Se muestra en la figura la representación gráfica de la ontología utilizando ONTOGRAF (figura 8).

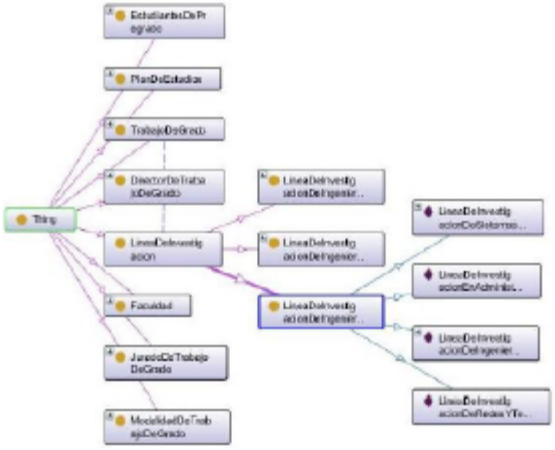

Fig. 8. Representación Gráfica de la Ontología de los Trabajos de Grado Utilizando ONTOGRAF Mostrando los Conceptos Principales e Instancias de la Subclase LineaDeInvestigacionDeIngenieriaDeSistemas

3. CONCLUSIONES

Para el modelo ontológico se desarrollan las fases de identificación, conceptualización, formalización, integración e implementación para poder cumplir con el propósito planteado en la investigación. Para el procesamiento del lenguaje natural se utiliza el conocimiento del dominio, desde la generación de textos, a la localización de ontologías, pasando por la recuperación de información, y la anotación lingüística basada en ontologías. Las ontologías se utilizan ya que es un mecanismo efectivo para representar conocimiento de manera formal, consensuada, legible por computadores como soporte para la web semántica, cuyo éxito, según sus impulsores, se materializará por la disposición a compartir ontologías que muestren comunidades y grupos en la web.

En la formalización del conocimiento se utiliza la representación del conocimiento del dominio de los trabajos de grado en una base de conocimiento, implementada como una Ontología de nivel intermedio, desarrollada utilizando el editor Protégé y siguiendo un proceso de ingeniería muy detallado descrito en la metodología de desarrollo de ontologías llamada Methontology.

En la integración se construye la Ontología reutilizando definiciones y conocimientos presentes en otras Ontologías, que serán integradas en la nueva Ontología en desarrollo. Esta integración debe realizarse durante todo el ciclo de vida, teniendo presente que es más significativa en la etapa de especificación y conceptualización que en la etapa de implementación. Para la implementación se utiliza un lenguaje formal computable OWL por ser uno de los lenguajes más importantes para la construcción de ontologías.

REFERENCIAS

Aguado de Cea, G., Álvarez de Mon y Rego, I., & Pareja Lora, A. (2002). Primeras aproximaciones a la anotación linguistico-ontológica de documentos de la web semántica: OntoTag. *Revista Iberoamericana de Inteligencia Artificial*, 37-49

Bateman, J. A. (1992). The Theoretical Status of Ontologies in Natural Language Processing. *Text Representation and Domain Modelling Ideas from Linguistics and AI,*. Berlin.

Echarte, F. (2006). Técnicas y lenguajes para la representación del conocimiento. *http://www.eslomas.com/index.php/archives/2006/12/14/tecnicas-y-lenguajes-para-larepresentacion-del-conocimiento/*.

Gannon, T. (2009). Framework for the analysis of the adaptability, extensibility, and scalability of semantic information integration and the context mediation approach. *Proceedings of the 42nd Hawaii International Conference on System Sciences*.

García Jiménez, A. (2004). Instrumentos de representación del conocimiento: tesauros versus ontologías. *Anales de documentación*, 79-95.

Gatius, M., & González, M. (s.f.). Un sistema de diálogo multilingue dirigido por la semántica. *Universidad Politécnica de Catalunya*.

Gatius, M., & Namsrai, T. (2012). A conversational system to assist the user when accessing web sources in the medical domain. *The Fifth International Conference on advances in computer-human interactions*, 160-164.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer.

Goñi Menoyo, J. M. (1998). *Arquitectura para representación del conocimiento léxico en sistemas de procesamiento de lenguaje natural.* Madrid: Universidad Politécnica de Madrid.

Gruber, T. (1993). Toward principles for the design of ontologies used for knowledge sharing. *Formal Ontology in comceptual analysis and knowledge representation*.

Izquierdo, R., Ferrández, O., Ferrández, S., Tomás, D., Vicedo, J. L., Martínez, P., & Suárez, A.

(s.f.). QALL-ME: Question Answering Learning technologies in a multilingual and multimodal environment. *Sociedad Española para el Procesamiento del Lenguaje Natural*.

Moya, D., & Macias, J. (s.f.). Auditoría de consultas para la web semántica orientada al usuario final. *Universidad autónoma de madrid*.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. MIT Press.

Onyshkevych, B. (1997). *An Ontological-Semantic framework for text analysis.* Pittsburgh: Carnegie Mellon University.

Puentes Velásquez, A. (2011). Ontologías: una técnica de representación de conocimiento. *Avances en Sistemas e Informática*, 211-216.

Puentes Velásquez, A. M. (2010). *Ontología para la búsqueda semántica de géneros de orquídeas Colombianas*, Cúcuta, UFPS.

Raskin, V., Hempelmann, C., & Triezenberg, K. (s.f.). Semanic Forensics: An application of ontological semantics to information assurance. *Purdue University*.

Vásquez Pérez, S. (2009). *Resolución de la ambiguedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN*. Alicante: Universidad de Alicante.

Vilares Ferro, J. (2005). *Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español.* La Coruña: Universidade da Coruña.