cenidet

Centro Nacional de Investigación y Desarrollo Tecnológico

Departamento de Ciencias Computacionales

TESIS DOCTORAL

Aplicación de Ontologías en la Construcción de una Base de Conocimiento para una Interfaz en Español hacia Bases de Datos

Presentada por

José Antonio Zárate Marceleño

Maestro en Ciencias Computacionales por el I. T. E. S. M. Campus Morelos.

Como requisito para la obtención del grado de:

Doctor en Ciencias en Ciencias de la Computación

Director de tesis:

Dr. Rodolfo A. Pazos Rangel

Co-Directores de tesis:

Dr. Alexander Gelbukh Kahn

Dr. Joaquín Pérez Ortega

Cuernavaca, Morelos, México. 11 de Diciembre de 2007

cenidet

Centro Nacional de Investigación y Desarrollo Tecnológico

Departamento de Ciencias Computacionales

TESIS DOCTORAL

Aplicación de Ontologías en la Construcción de una Base de Conocimiento para una Interfaz en Español hacia Bases de Datos

Presentada por

José Antonio Zárate Marceleño

Maestro en Ciencias Computacionales por el I. T. E. S. M. Campus Morelos.

Como requisito para la obtención del grado de:

Doctor en Ciencias en Ciencias de la Computación

Director de tesis:

Dr. Rodolfo A. Pazos Rangel

Co-Directores de tesis:

Dr. Alexander Gelbukh Kahn

Dr. Joaquín Pérez Ortega

Jurado:

Dra. Azucena Montes Rendón Presidente	Dr. Grigori Sidorov Vocal
Dr. Rodolfo A. Pazos Rangel Secretario	Dr. Víctor Jesús Sosa Sosa Vocal suplente

Cuernavaca, Morelos, México. 11 de Diciembre de 2007

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

SIGLARIO

Siglas	Significado
ILNBD	Interfaz de lenguaje natural hacia bases de datos.
HTML	Hypertext Markup Language (Lenguaje de Marcado de Hipertexto).
OWL	Web Ontology Language (Lenguaje de Ontologías de Web).
PLN	Procesamiento de lenguaje natural.
RDF	Resource Description Framework (Marco de Descripción de Recursos).
URI	Uniform Resource Identifier (Identificador Uniforme de Recursos).
URL	Uniform Resource Locator (Localizador Uniforme de Recursos).
XML	eXtensible Markup Language (Lenguaje de Marcado Extensible).
ELF	English Language Front-End (ILNBD para el lenguaje inglés).
EQ	English Query (un componente de MS-SQL Server).
GPL	General Public License (Licencia Pública General).

CAPÍTULO 1
Introducción

En este capítulo se explica el contexto del proyecto de tesis, el problema que se aborda, el objetivo y los alcances del mismo, además de la organización general del documento.

1.1 Antecedentes

El rápido crecimiento de la Internet (WWW) está formando una sociedad donde la demanda de servicios de almacenamiento, organización, acceso y análisis de la información va en constante aumento. Los sistemas de pregunta-respuesta (QA por sus siglas en inglés) han llegado a ser una alternativa real a los sistemas de recuperación de información tradicionales, debido a su capacidad de proporcionar respuestas concretas a consultas hechas por el usuario en lenguaje natural. Este hecho, junto con la inclusión de la evaluación de sistemas QA en la Conferencia de Recuperación de Texto (TREC por sus siglas en inglés) en 1999 [3], y recientemente en el Foro de Evaluación a través del Lenguaje (CLEF) [19], muestra el grado de interés que existe en este campo de la investigación. Actualmente, los sistemas de QA se han enfocado a las consultas de hechos [36] que requieren como respuesta entidades nominales (fecha, cantidad, nombre propio, localidad, etcétera).

Estas entidades nominales se encuentran normalmente en bases de datos, lo cual cae en el campo de las Interfaces de Lenguaje Natural hacia Bases de Datos (ILNBDs). Este tema ha despertado interés desde la década de los años 70s, pero aún no han sido completamente solucionados los problemas implícitos en su desarrollo. La mayoría de las ILNBDs no son completamente Interfaces de Lenguaje Natural, ya que sólo el componente de consultas está definido para aceptar un lenguaje restringido al contexto de la base de datos, aunque algunas interfaces permiten también actualización restringida de los datos [1].

Encuestas realizadas a estudiantes de maestría de dos universidades, mostraron que más del 90% de ellos no conocían las ILNBDs ni ninguna interfaz del mismo tipo. Lo anterior es un ejemplo de la **poca difusión de las ILNBDs y, por lo tanto, lo limitado de su utilización, debido a la complejidad para adaptarlas a las necesidades de los usuarios finales**. A pesar de que las interfaces comerciales cuentan con una estrategia de adaptación automática al contexto de la base de datos, aún es necesario modificar la configuración resultante del proceso anterior, por medio de la intervención de un ente externo (administrador de la base de datos, alguien del equipo de desarrollo de la ILNBD, etcétera).

La tarea a realizar por el encargado de personalizar la interfaz no es muy clara, ya que para minimizar su trabajo, la ILNBD le solicita un poco de información complementaria a la configuración obtenida automáticamente en un análisis anterior. Esta información complementaria se limita a definir palabras en el diccionario de la interfaz y agregar información de cómo se relacionan las palabras.

Analizando la información disponible de los procesos de configuración de las ILNBDs más representativas, se encontraron ciertas deficiencias en cuanto al conocimiento que un personalizador necesita proporcionar a una ILNBD para que ésta funcione adecuadamente. En esta tesis se propone una modificación a la arquitectura básica de una ILNBD (figura 2-1) que se caracteriza por usar como

base de conocimiento una ontología, la cual, sin estar en contraposición al proceso de autoconfiguración y a los medios que proporcionan las ILNBDs comerciales para ajustar la configuración, ofrece como novedades la incorporación de principios de reuso, explicitez de la base de conocimiento, clasificación de los tipos de consulta, generalidad, simplicidad y los medios para explotar la ontología.

Se hicieron pruebas comparativas entre el proceso de configuración de la ILNBD comercial más disponible (English Query, un componente de Microsoft SQL Server [20]) contra un proceso de configuración propuesto basado en una ontología con grupos de estudiantes de maestría, debido a la capacidad de análisis necesaria para llevar a cabo el proceso de configuración, siendo los resultados favorables a lo propuesto.

1.2 Descripción del problema

Una ILNBD es un programa que recibe las consultas que le hace un usuario en lenguaje natural, y las traduce a un lenguaje de consulta, siendo el más común SQL. Se busca que el usuario no tenga que aprender SQL u otro lenguaje formal para acceder a los datos de una manera natural, tal como si se los estuviera pidiendo a otra persona. Para lograr lo anterior, es necesario que la ILNBD tenga cierta información morfológica, sintáctica y semántica, que le permita validar y "entender" la consulta del usuario.

La información morfológica es factible de ser incluida en la ILNBD a través de un lexicón (diccionario) predefinido. La información sintáctica que valida que un verbo puede ir después de un sustantivo en una oración, se puede introducir en una ILNBD a través de una gramática predefinida, pero la información semántica en general, el significado de las palabras y de las oraciones, es prácticamente imposible de prefijar en la ILNBD, ya que depende muchas veces del contexto en que se hace la consulta.

Un ejemplo de lo anterior se da al definir que el concepto *empleado* lo une la relación *actor* con el verbo *trabajar*. Si es difícil que una ILNBD obtenga la información anterior de un análisis no supervisado de la base de datos, lo es más que incluya información muy específica como la siguiente: el concepto *exportador* está unido por la relación *medio_transporte* con el concepto *buque*. Lo anterior implica la necesidad de un *ingeniero del conocimiento* o en nuestro caso, un personalizador de la interfaz que revise la información predefinida de una ILNBD, la adapte en caso de ser necesario y le agregue aquella información particular del contexto, que es muy difícil que los desarrolladores de la ILNBD le hayan incluido, o que ésta pueda obtener automáticamente de un análisis de la base de datos y de los metadatos.

Uno de los principales motivos por el cual es muy difícil configurar automáticamente la información semántica de una ILNBD, se da por las variantes del lenguaje. Un ejemplo de dichas variantes se puede apreciar en la consulta "dame los nombres y apellidos de los empleados", la cual sería fácilmente contestada por la ILNBD si en la base de datos una tabla se llamara *empleado*, y

dos de sus columnas se llamaran *nombre* y *apellido*. Si la consulta anterior se replanteara como “dame los nombres completos de los trabajadores”, la ILNBD tendría que tener en su base de conocimiento que trabajador es sinónimo de *empleado* (una variante), y que *nombre* y *apellido* son parte del nombre completo de una persona en general (otra variante). Esta información faltante tendría que ser proporcionada por un *personalizador*, que podría ser el administrador de la base de datos, algún miembro del equipo que la diseñó, o en principio cualquier profesional de la computación.

Existen diferentes ILNBDs, y cada una posee diferentes métodos de personalización: en Masque [2] y NATLIN [30] es necesario modificar el código (en lenguaje Prolog) para adaptar la interfaz a una determinada base de datos; en ELF [7] y English Query [20], se tienen interfaces de configuración, en donde se supone que el configurador tiene amplios conocimientos de gramática. El enfoque que se presenta en esta tesis se basa en el hecho de que el conocimiento tiene que estar organizado para poderse utilizar, pero esta organización debe ser lo más sencilla posible, y a su vez, lo suficientemente flexible para expresar el conocimiento que necesita la ILNBD [39].

Además de organizar el conocimiento, se propone una metodología que permita hacer más flexible la ILNBD, no para que pueda adaptarse a diferentes bases de datos (lo cual en cierta manera ha sido alcanzado) sino para que la ILNBD pueda adaptarse a los diferentes problemas relacionados con la forma en que se hacen las consultas a la base de datos.

Resumiendo, el problema que se estudió fue: **el *proceso de configuración de las interfaces en lenguaje natural hacia bases de datos que limita su adaptación a los diferentes tipos de consultas del usuario final y su portabilidad de dominio.***

1.3 Objetivo

El objetivo general de este trabajo fue desarrollar una metodología para configurar una ILNBD vía una ontología, que fuera más flexible y aceptable por los encargados de personalizarla, que las propuestas por otras ILNBDs. Esta mejora en el proceso de configuración de las ILNBDs, permitirá que a los encargados de personalizarla les sea más fácil proporcionar la información necesaria para que la interfaz conteste las consultas que le hacen los usuarios finales, ya que la complejidad del proceso influye en la poca aceptación y difusión de este tipo de interfaces. La mejora propuesta consiste en la adaptación de una ILNBD hacia una determinada base de datos, a través del uso de una ontología y los mecanismos de explotación de la misma. Además del objetivo general, se plantearon los siguientes objetivos particulares:

- Desarrollar una metodología para configurar una ILNBD vía una ontología que cubriera algunas deficiencias encontradas en otras ILNBDs.
- Diseñar una ontología que cumpliera las expectativas de ser flexible, fácil de entender y que sirviera como soporte a la metodología propuesta.

- Desarrollar las herramientas necesarias para poder crear y explotar la ontología.
- Desarrollar una metodología que permitiese comparar la propuesta de configurar una ILNBD vía una ontología contra el procedimiento usado actualmente por las interfaces comerciales, diferente de las evaluaciones tradicionales de este tipo de interfaces, las cuales se enfocan a medir el número de respuestas correctas contra el número de consultas hechas a la ILNBD.

1.4 Hipótesis

H1. *El uso de una ontología para representar el conocimiento lingüístico, reduce el esfuerzo de configurar una ILNBD para diferentes tipos de consultas.*

H2. *El uso de una ontología para representar el conocimiento lingüístico, reduce el esfuerzo de portar una ILNBD para diferentes dominios.*

1.5 Alcances y limitaciones

Los alcances de este trabajo son los siguientes:

- Diseñar una ontología que cumpla con los requerimientos de simplicidad, reuso, extensibilidad y que cumpla con la misión de servir de enlace entre los elementos de la base de datos y los elementos que constituyen la consulta del usuario final.
- Implementar un mecanismo de aprendizaje de la ontología que permita tener una ontología prellenada y reusable [40].
- Implementar un módulo que explote la información de la ontología, que permita ser usado para traducir una consulta a SQL.

Las limitaciones de este trabajo son las siguientes:

- Las pruebas comparativas se realizaron con English Query, debido a que se considera que su mecanismo de configuración es el más completo actualmente.

- No incluye una solución a los problemas lingüísticos que existen en el español, tales como la elipsis o la anáfora, por citar algunos.
- Las pruebas se efectuaron con bases de datos relacionales.
- La ontología sólo fue usada como base de conocimiento y no para cualquier otro uso que pudiera tener.
- No se incluye el desarrollo de un analizador sintáctico, analizador semántico, analizador pragmático o del discurso.

1.6 Aportaciones

- Una metodología de configuración más detallada que las definidas para otras ILNBD.
- Una ontología que cubre aspectos léxicos y del modelo relacional que sirviera como puente entre una consulta en lenguaje natural y su equivalente en SQL.
- Una primera aproximación a la evaluación de ILNBDs desde el punto de vista del configurador.
- Un mecanismo de aprendizaje que incorpora el uso de un análisis semántico para encontrar los ejemplares de la ontología.

1.7 Organización del documento

Este documento de tesis se organiza de la siguiente manera:

Capítulo 2. Se presenta el marco teórico donde se incluyen los conceptos relacionados con la tesis así como una descripción y comparación de los trabajos relacionados.

Capítulo 3. Se describe la propuesta de implementar la base de conocimiento de una ILNBD usando ontologías.

Capítulo 4. Se describen las pruebas que sirven para validar las hipótesis planteadas para la propuesta del uso de ontologías.

Capítulo 5. Se presentan las conclusiones obtenidas, trabajos futuros y aportaciones de este trabajo.

CAPÍTULO 2
Marco teórico y estado del arte

En este capítulo se explican los conceptos en los cuales se fundamenta este trabajo, abarcando una introducción al procesamiento de lenguaje natural y ontologías, junto con una revisión de los trabajos relacionados.

2.1 Procesamiento de lenguaje natural

El **procesamiento de lenguaje natural (PLN)** es el estudio de mecanismos que permitan la comunicación entre una computadora y una persona, por medio del lenguaje natural. El procesamiento de lenguaje natural es un campo creado a partir de la unión de la ciencia computacional con la lingüística, y se enfoca en los problemas del modelado de lenguaje natural en una computadora. Los lenguajes naturales tales como el inglés, el francés, el español, etc., son los que las personas emplean para comunicarse entre sí. Un **lenguaje** puede definirse simplemente como un conjunto de cadenas de caracteres sin referencia a ningún mundo por describir ni a ninguna tarea por desarrollar.

La definición de lenguaje natural proviene de la teoría lingüística y fisiológica de que la función del lenguaje, tal como se manifiesta cuando hablamos, es enteramente natural, es decir, que nuestro aparato vocal está hecho para hablar como nuestro oído está hecho para oír. Un investigador de apellido Broca descubrió que la facultad de hablar está localizada en la tercera circunvolución frontal izquierda del cerebro, demostrando con esto el carácter natural del lenguaje [15]. En la inteligencia artificial el uso del concepto de "lenguaje natural" es utilizado para distinguir este tipo de lenguajes (por ejemplo: el español) de los lenguajes *"artificiales"* tales como C, Pascal, COBOL, Java, etc.

La mayor parte de la comunicación lingüística humana se produce, con mucho, de forma oral, aunque con el advenimiento de los medios electrónicos, la escrita también se ha incrementado notablemente. El procesamiento del lenguaje escrito es diferente que el del hablado, ya que en el escrito se delimita el alcance, y el texto en sí, puede dar una idea más completa de lo que se quiere expresar; mientras que el lenguaje hablado implica conocer elementos que no necesariamente están dentro de la conversación. Por lo tanto, es positivo dividir la tarea global del procesamiento de lenguaje en dos partes:

- **Procesamiento de lenguaje escrito**. Éste usa conocimiento léxico, sintáctico, semántico, contextual y pragmático sobre el lenguaje y el mundo real.
- **Procesamiento de lenguaje oral.** Éste es semejante al procesamiento anterior, excepto que se necesita adicionalmente conocimiento sobre fonología, e información adicional para manejar las posibles ambigüedades que pudieran surgir en el habla [17].

2.2 Enfoques de procesamiento de lenguaje natural

- **PLN General.** Su objetivo es modelar el lenguaje, a partir del uso que el ser humano le da; es decir, es una simulación por computadora. Algo que se ha aprendido de este enfoque, es que se requiere una gran cantidad de conocimiento del mundo real.

- **PLN Aplicado**. En éste no es importante realizar simulaciones cognoscitivas sino llegar a productos aplicados a la comunicación conversacional a través de lenguaje natural entre usuarios y computadoras [17].

2.3 Fases del procesamiento de lenguaje natural

- **Fonología.** Es el estudio de la estructura sonora del lenguaje. Los sonidos son organizados en una estructura de contrastes y analizados en términos de fonemas. Un fonema es la unidad mínima del sistema de sonidos de un lenguaje.
- **Morfología.** Es el estudio de las estructuras que permiten la formación de las palabras, separando de éstas sus elementos básicos o morfemas, a partir de un proceso de lematización (obtención de la raíz de la palabra).
- **Análisis sintáctico (parsing).** Es el estudio de las estructuras que combinadas forman oraciones. Las estructuras sintácticas (o construcciones) son analizadas en secuencias de categorías sintácticas (o clases). El orden de dichas estructuras es establecido sobre la base de relaciones sintácticas.
- **Análisis semántico.** Es el estudio del significado de las construcciones de un lenguaje. Se hace una correspondencia entre las estructuras sintácticas y los objetos del dominio de la tarea.
- **Integración del discurso.** Es el proceso de asignar un significado a una frase individual dependiendo de las frases precedentes, e influyendo en el significado de las frases posteriores (referencia, anáfora y catáfora).
- **Pragmática.** Es el estudio del uso comunicativo del lenguaje en un contexto, particularmente la estructura de la conversación y el diálogo. La estructura que representa se reinterpreta para determinar su significado (elipsis) [17].

2.4 Arquitectura de un sistema de procesamiento de lenguaje natural

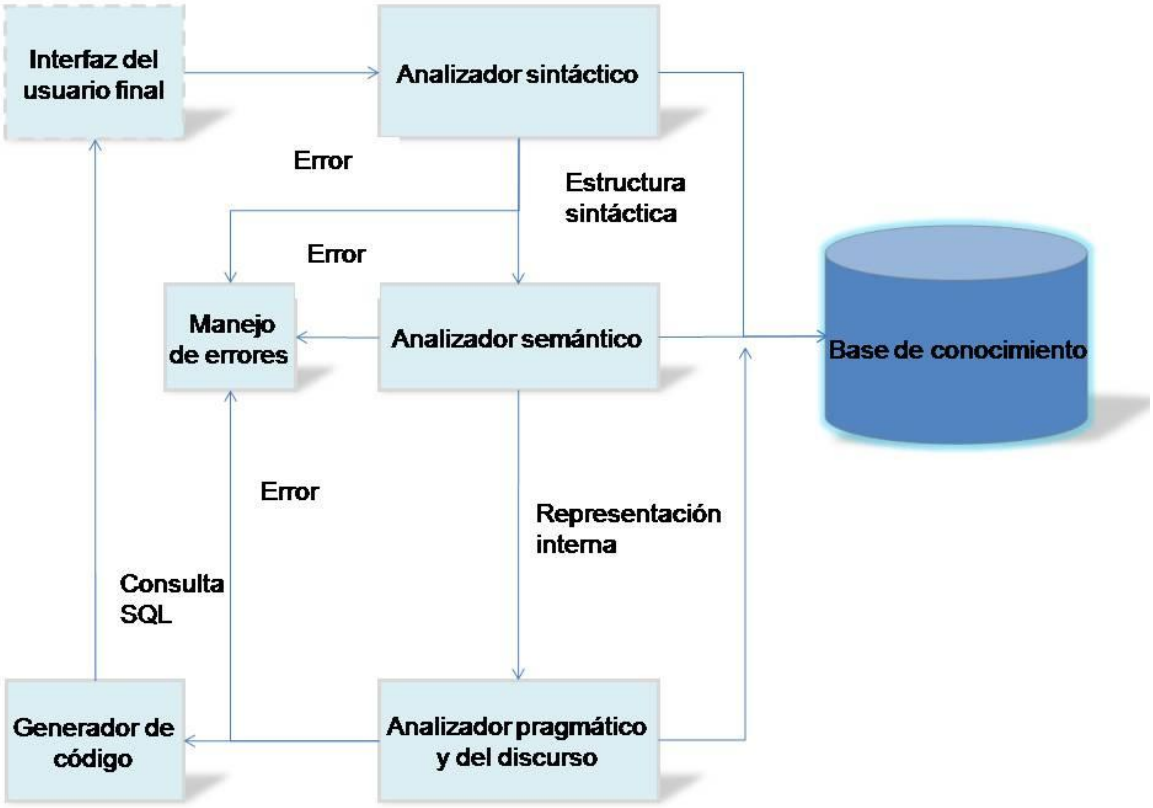

Figura 2-1 Arquitectura básica de un sistema de PLN.

La arquitectura mostrada en la figura 2-1 funciona de la siguiente manera:

Se da una oración en lenguaje natural, ya sea por medio de un teclado, por un archivo o por medio de una interfaz de voz; se pasa esta entrada a un analizador sintáctico para que determine si la oración cumple las reglas definidas por la gramática, generando, en caso de que la entrada sea válida, una estructura sintáctica que represente las construcciones y clases sintácticas de la oración de entrada.

La estructura sintáctica anterior es procesada por el analizador semántico para asignarle un significado a partir de algún formalismo de representación del conocimiento, buscando sobre todo eliminar las ambigüedades presentes en la oración; en caso de que no cumpla las reglas sintácticas o semánticas del lenguaje, se notificará del error al usuario.

Los dos procesos anteriores, análisis sintáctico y semántico, se apoyan en el uso de una estructura llamada base de conocimiento, que contiene el significado de las palabras y sus estructuras sintácticas. Esta base proporciona un gran rango de información acerca de las palabras, pronunciación, morfología, sintaxis, relaciones semánticas y léxicas, para definir el sentido de las palabras o cadenas de palabras, según sea el caso, para que el sistema de PLN tenga el conocimiento no sólo del significado de las palabras sino del mundo mismo.

Por último, a partir de que se comprende qué es lo que significa realmente la oración de entrada, un módulo traductor transforma la estructura generada por el analizador semántico a su forma equivalente en un lenguaje formal, la cual pueda ser más fácilmente procesada por una computadora o cualquier otro dispositivo.

La comprensión es el proceso de mapear un mensaje de entrada a otra representación más útil. Para entender esto, es necesario que pensemos en el lenguaje como un par (lenguaje fuente, representación en el lenguaje de destino), junto con una correspondencia entre los elementos de uno y del otro. La representación destino se elige de acuerdo con la tarea que se quiere desarrollar.

2.5 Clasificación de los sistemas de procesamiento de lenguaje natural

Se han propuesto diferentes esquemas de clasificación para los sistemas de PLN, siendo uno de ellos el propuesto por Hendrix y Sacerdote [15], el cual se basa en el nivel de complejidad que alcanza el sistema. Se propusieron tres niveles de clasificación, basados en la cantidad de conocimiento del mundo exterior que el sistema posee:

- **Tipo A: Modelos sin conocimiento del mundo**. Utilizan únicamente la sintaxis y la semántica de las oraciones para entender el diálogo. Para entender el significado de una unidad sintáctica más grande, se basan en el significado de las partes que la componen.
- **Tipo B: Sistemas que utilizan explícitamente modelos del mundo**. Estos sistemas contienen conocimiento de un dominio, y para poder entender las entradas del usuario, es necesario conocer el contexto y las restricciones físicas involucradas.
- **Tipo C: Sistemas que incluyen modelos del mundo humano**. La investigación en lingüística computacional ha revelado poco a poco la importancia de las creencias, metas y los planes para lograr una buena comunicación entre humanos. Este hecho ha sugerido la incorporación de este tipo de conocimiento en los sistemas de PLN, aunque se ha visto que es una de las metas más difíciles de alcanzar en la investigación en inteligencia artificial. No se conoce ningún sistema de PLN que haya alcanzado este nivel de

sofisticación, por ello, un buen número de grupos de investigación está estudiando los elementos del problema.

Existe otra clasificación propuesta en relación al uso que se le dará al sistema de PLN:

- **Sistemas generadores de lenguaje natural.** Se utilizan para generar una comunicación basada en los conocimientos del lenguaje y del destinatario de la comunicación. Algunos ejemplos de su aplicación son la generación automatizada de memorándums, o el servicio de atención por teléfono.
- **Sistemas para la comprensión de lenguaje natural.** Considerando al lenguaje como un par (lenguaje fuente, representación destino), se define la comprensión de lenguaje natural como el proceso de cambiar una determinada forma de entrada a otra representación más útil [5].

2.6 Interfaces de lenguaje natural hacia bases de datos

Una aplicación particular de los sistemas de PLN son las Interfaces de Lenguaje Natural hacia Bases de Datos (ILNBDs). Esta área de PLN, en la que se centra esta investigación, trata de dar un nuevo enfoque al problema de interfaces "amigables" a bases de datos, ya que, siendo las bases de datos el principal recurso a consultar, es deseable que el usuario pueda, de la manera más sencilla, obtener la información que necesita.

La mayoría de las ILNBDs sólo permiten consultas sobre los datos, aunque existen algunos sistemas que también permiten la actualización. Las principales razones para utilizar las ILNBDs son las siguientes [1]:

- Eliminan lo artificial de los lenguajes formales.
- Son mejores para tipos de consultas complejas y que involucren varias relaciones.
- Permiten hacer suposiciones basadas en el contexto.
- Disminuyen notablemente el tiempo de aprendizaje del usuario.
- Permiten que prácticamente cualquier usuario pueda acceder a los datos fácilmente.

Por otra parte, los inconvenientes encontrados en estos sistemas son los siguientes:

- La cobertura lingüística no es obvia.
- Pueden llegar a ser un medio de comunicación difícil.

- Su configuración puede ser lenta y tediosa (aunque generalmente después de la primera vez, su mantenimiento debe ser mínimo).
- A veces el usuario supone habilidades que la interfaz no posee.

En un análisis más específico de las ILNBDs, se encontró que una de sus grandes desventajas es que los usuarios suponen que la computadora es más inteligente de lo que realmente es. En un experimento con el sistema SUNDIAL (con una interfaz de voz), se llegó a la conclusión de que la gente tiende a simplificar su manera de hablar al momento de darse cuenta que está hablándole a una máquina [31].

2.7 Problemas actuales de las interfaces de lenguaje natural hacia bases de datos

A pesar de que existen ILNBDs que ofrecen portabilidad hacia múltiples tipos de sistemas (sistemas expertos, manejadores de bases de datos, etc.), múltiples sistemas operativos y múltiples sistemas manejadores de bases de datos [31] y que además, fueron diseñadas de una manera modular que les permite cierta versatilidad [14], la portabilidad de un dominio de conocimiento a otro (es decir, la capacidad de que la misma interfaz pueda trabajar con diferentes bases de datos), no ha sido totalmente alcanzada. Esto se debe a que aquellas ILNBDs que aseguran que son fácilmente trasladables de un dominio a otro, se basan en el hecho de que se puede acceder a su base de conocimientos y modificarla para el nuevo dominio de conocimiento. La solución anterior no toma en cuenta que para poder modificar la base de conocimiento, ésta debe ser construida en forma tal que se le hagan únicamente los cambios necesarios para adaptarse a un nuevo contexto y que además, si se dispone de otro "conocimiento" definido por terceros, se pueda reutilizar para construir la base de conocimiento para el nuevo dominio.

Lo anterior es un punto importante, dado que necesariamente "alguien" (llámese personalizador, administrador de la base de datos o el diseñador de la misma) debe configurar la ILNBD, cambiando las reglas gramaticales de acuerdo a la naturaleza de la base de datos y agregando palabras al diccionario y las relaciones semánticas entre estas palabras, de acuerdo al nuevo contexto.

2.8 Técnicas de configuración de las interfaces de lenguaje natural hacia bases de datos

ELF [7] realiza un análisis automático tanto de los metadatos como de los datos, para generar una configuración de la interfaz para una base de datos determinada. Este análisis se basa en un diccionario, un lexicón, la descripción de las columnas y en un conjunto de atributos predefinidos. A partir de esta información, ELF determina los sustantivos relacionados con columnas y tablas, y los verbos que identifican las relaciones que hay entre tablas y columnas. Además, ELF permite modificar el lexicón, el cual contiene la información recolectada durante el análisis, en el caso de que haya habido alguna confusión. También permite revisar el

diccionario (Moby dictionary), un diccionario predefinido de 17,000 entradas, varias de las cuales son sinónimas entre sí. Una característica interesante es que permite la extensión de la interfaz por medio de guiones (scripts) en Visual Basic, macros y llamados a funciones.

English Query [17] realiza también un análisis muy semejante al de ELF, pero no está disponible el lexicón y su análisis es más limitado. Éste se restringe a relacionar columnas de la base de datos con palabras, predefinir ciertas relaciones como *tener* (muy genérica, ya que sólo establece que una entidad tiene columnas) e *identificador único* (conceptos que identifican a una tabla). Además cuenta con un diccionario de sinónimos, y permite establecer ciertas relaciones basadas en frases, por ejemplo: “los proveedores surten pedidos”. Permite definir relaciones temporales entre los conceptos de la base de datos, relaciones de hiperonimia-hiponimia entre relaciones, y agregar cierta funcionalidad a la interfaz por medio de enlaces entre frases y llamados a funciones externas (característica compartida con ELF). La última versión está integrada con Visual Studio 6.0, permitiendo con esto definir relaciones entre conceptos que representan entidades, de una manera gráfica, muy semejante a un diagrama Entidad-Relación (figura 2-2). También proporciona información que usó la interfaz para contestar una consulta, y además, cuenta con un asistente que le permite a la persona que configura retroalimentar a la interfaz con la información que ésta necesita, cuando falla al dar una respuesta. Esta retroalimentación consta de una serie de cuestionarios que deben llenarse con información adicional a la consulta.

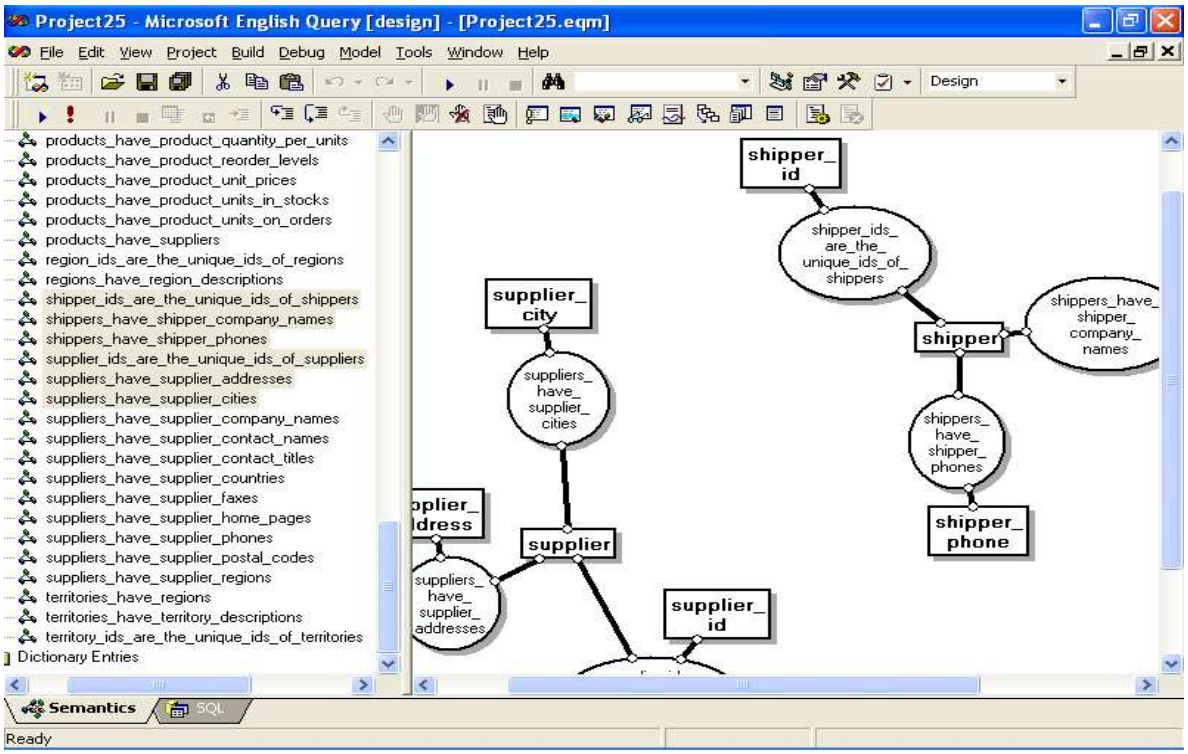

Figura 2-2 Ambiente de configuración de English Query.

Inbase [32], una ILNBD desarrollada en el Instituto Ruso de Investigación en Inteligencia Artificial, basa su tecnología en la separación del conocimiento acerca de patrones semánticos que son usados para consultar la base de datos y el conocimiento del problema de una base de datos en particular. Esto permite que se ajusten rápidamente las capacidades del componente de análisis de lenguaje natural hacia la base de datos que se va a consultar. Para contestar las consultas se necesita un modelo del dominio (DM), el cual se obtiene en parte por un análisis de la base de datos, y en parte por información que proporciona un diseñador. El modelo del dominio está formalizado en SNOOP [33], un derivado del diseño orientado a objetos.

2.9 Revisión de trabajos relacionados

A pesar de que ELF se considera como una de las mejores ILNBDs disponibles [28] y que según su documentación, sólo es necesario un mínimo de esfuerzo extra para afinar su preconfiguración, se encontraron algunos problemas con esta manera de configurar: sus atributos predefinidas no están organizados, ya que en muchos casos mezclan elementos relacionados con el análisis sintáctico de las consultas con información semántica, y tampoco es factible agregar nuevos atributos. En la documentación de ELF se menciona que el análisis automático que realiza detecta relaciones de sinonimia, pero no aclara si la interfaz puede manejar otro tipo de relaciones (antonimia, meronimia, etc.) y de qué manera se le podrían definir nuevas relaciones. Hay muchos conceptos de sintaxis y semántica que debe manejar el configurador de la interfaz al mismo tiempo, dificultándole con esto su trabajo.

Los problemas encontrados en English Query se describen a continuación: Aunque se le pueden agregar palabras al diccionario, es algo confusa la manera en que esto se hace y el uso que se le da a estas nuevas palabras. También se pueden agregar sinónimos y nuevas relaciones entre conceptos, aunque de una manera muy rígida, ya que obedece a ciertos patrones de oraciones (frases nominales, verbales, adjetivales, adverbiales, comando y preposición), y no queda claro cuál es la diferencia entre definir la relación usando uno u otro patrón. Además, las relaciones que encuentra English Query son muy genéricas y no muy útiles. El uso del asistente de retroalimentación no es muy intuitivo, ya que, para consultas similares que no son correctamente contestadas por la interfaz, puede ser necesario proporcionar diferente información, para que English Query las pueda contestar correctamente.

En el caso de Inbase [32], el prototipo de demostración mostrado en línea no se puede configurar y sus respuestas no son muy confiables, ya que no sabe distinguir variantes de una consulta (por ejemplo: “Cuál es el empleado con mayor sueldo” y “Cuál es la edad del empleado con mayor sueldo”). Se desconoce si este problema sólo ocurre en la versión en inglés o si las versiones en otros idiomas son más precisas. No se encontró la descripción del proceso de configuración, aunque en algunas referencias del proyecto se menciona que utilizan KL-ONE [38], uno de los lenguajes de representación de conocimiento más estables.

Debido a lo anterior, no fue posible evaluar la complejidad o sencillez del proceso de configuración de esta ILNBD (y otras tantas como PRECISE [25] por las mismas razones).

En base a la revisión de las ILNBDs descritas anteriormente, se concluyó que actualmente existen problemas al personalizar una base de conocimiento de una interfaz en lenguaje natural hacia bases de datos. Es necesario que exista una metodología de diseño que permita hacer esta configuración de una manera semejante a como se construye software, es decir, que se haga uso de componentes genéricos y que aquellos detalles particulares de la aplicación sean los únicos que necesiten ser modificados o agregados. Además, una de las tendencias más importantes es la de compartir información, con el fin de que se puedan disminuir los costos que implica construir sistemas basados en conocimiento.

2.10 Ontologías

Una ontología se define como una especificación explícita de una conceptualización [13]. Es una descripción formal de conceptos pertenecientes a un dominio, así como las relaciones existentes entre ellos. El estudio de las ontologías nace de la filosofía y proviene de las etimologías griegas *"ontos"* (el ser) y *"logos"* (estudio), lo cual significaría el estudio del ser.

Una ontología es una especificación explícita de conceptos que definen una vista abstracta, simplificando las características del mundo que necesitamos representar con un determinado propósito. En esta conceptualización se definen los objetos y otras entidades que se supone que existen en un área de interés y las relaciones entre ellas; además, se define un vocabulario de términos y especificación del sentido de los conceptos definidos en la ontología.

Por definición del A*merican Heritage Dictionary*, ontología es "la rama de la metafísica asociada con la naturaleza del ser". En cambio, la comunidad de inteligencia artificial la define como el conjunto de conceptos o términos que pueden ser usados para describir algún área del conocimiento o construir una representación de ella. Una ontología puede consistir de conceptos de muy alto nivel, organizados sobre una base de conocimientos [21].

En el enfoque computacional, las ontologías abordan el desarrollo de un esquema conceptual sobre un dominio determinado [10], donde se clasifican conceptos y las interrelaciones existentes entre ellos. La definición de *clases* generalmente es el tema central de la mayoría de las ontologías.

Las clases describen los *conceptos* dentro de un dominio. Por ejemplo, consideremos una conceptualización en que la clase *vehículos* representa todas las marcas de vehículos. En este ejemplo, las *subclases* serían cada una de las marcas. Se llama *ejemplar* de una clase cuando se habla de un elemento específico y particular que pertenece a una clase. Siguiendo el mismo ejemplo, el Volkswagen rojo con placa X sería un ejemplar que pertenece a la clase

Volkswagen, subclase de la clase automóvil, y ésta a su vez subclase de la clase vehículos.

Dos de las relaciones más utilizadas en una ontología para interrelacionar conceptos son: la relación de generalización o especialización, llamada *hiperonimia* (relación *es un*), y la relación de pertenencia, conocida como *meronimia* (relación *parte de* o *tiene*). Un ejemplo de hiperonimia se ilustra en la figura 2-3.

El uso de ontologías cada vez está tomando mayor auge, gracias a que por medio de ellas se puede hacer referencia a una misma conceptualización, logrando unificar y formalizar los conceptos de un dominio, clasificarlos y darles un orden. Existe una serie de consideraciones necesarias a tomar en cuenta al momento de desarrollar una ontología; aunque no existe una manera única y correcta de modelar un dominio:

- El diseño de una ontología es un proceso iterativo.
- Los conceptos de una ontología, lógicos o físicos, así como las relaciones entre ellos, deben ser relevantes al dominio de interés.

Con respecto al primer punto, hay que considerar que existen distintos puntos de vista y consideraciones hacia un mismo tema, por lo que es lógico pensar que un mismo dominio puede ser conceptualizado de diferentes maneras, sin implicar que alguna de ellas esté bien o mal, porque son maneras distintas de interpretar el mundo.

El conocimiento definido en una ontología se puede ver como "metainformación"; es decir, información semánticamente "rica". El grado de formalidad de esta especificación va desde muy informal, seminformal, formal hasta rigurosamente formal. Las ontologías tienen los siguientes componentes que sirven para representar el conocimiento de algún dominio [10]:

- **Conceptos**. Son las ideas básicas que se intentan formalizar. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, etc.

- **Relaciones**. Representan la interacción y enlace entre los conceptos del dominio y suelen formar la taxonomía del dominio. Por ejemplo: *subclase-de*, *parte-de*, *parte-exhaustiva-de*, *conectado-a*, etc.

- ***Funciones***. Son un tipo concreto de relación donde se identifica un elemento mediante el cálculo de una función que considera varios elementos de la ontología (por ejemplo: *categorizar-clase*, *asignar-fecha*, etc.).

- **Ejemplares**. Se utilizan para representar objetos determinados de un concepto.

- **Axiomas**. Son teoremas que deben cumplir los elementos de la ontología. Por ejemplo: *si A es subclase de C y B subclase de A, entonces C no debe ser subclase de B*, etc.

Las ontologías han servido, sobre todo, para conformar una vista común de la representación de teorías de dominios particulares y para poder comunicar y hacer interoperables diferentes aplicaciones. Las ontologías han sido usadas principalmente en los siguientes campos:

a) Comunicación.

- Para modelos normativos, creando la semántica de un sistema y el modelo para extenderlo y transformarlo entre diferentes contextos.
- Como una red de relaciones.
- Para lograr consistencia y disminuir la ambigüedad.
- Integración de diferentes perspectivas de usuarios.

b) Interoperabilidad (ontologías como una interlingua).

- Interna: bajo el control de una unidad organizacional.
- Externa: abierta a cualquier tipo de organización.
- Integrada entre dominios (por ejemplo: modelar conocimiento de ingeniería química).
- Integrada entre herramientas (es útil cuando cada herramienta usa su propio lenguaje de representación y manejo de la ontología).

c) Ingeniería de sistemas.

- Especificación de requerimientos, al ayudar a encontrar posibles ambigüedades.
- Reutilización del conocimiento para la construcción de otros sistemas.

En principio, se puede usar cualquier lenguaje de programación para implementar las ontologías, pero a veces los lenguajes más comunes (C++, Java, Delphi) carecen de expresividad para escribir lo que se desea decir. Algunas de las primitivas que los lenguajes deben añadir para mejorar la representación de ontologías son las siguientes:

- Constructores para agregados, múltiples jerarquías clase-subclase, reglas y axiomas.

- Mecanismos de diseño basado en módulos para poder escribir diferentes ontologías y sus interrelaciones.
- La posibilidad de tener una visión a un metanivel [21].

Un ejemplo de una ontología se muestra en la figura 2-3.

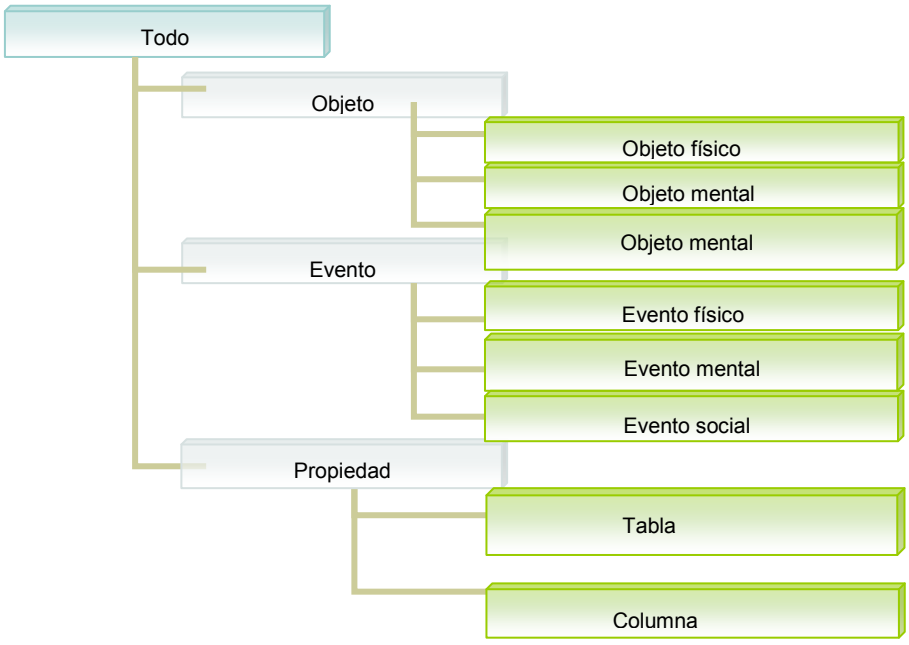

Figura 2-3 Ejemplo de una ontología.

2.11 Lenguajes para representar de manera formal una ontología

Una ontología formal se compone de un vocabulario controlado, expresado en un lenguaje de representación ontológica. Se ha desarrollado una serie de lenguajes de descripción, algunos de los cuales se describen a continuación:

Resource Description Framework (RDF) [27] es una recomendación del W3C (World Wide Web Consortium) para la representación de metadatos dentro de la Web. Es una especificación que proporciona una manera sencilla de intercambiar conocimiento por medio de la Internet, tomando como base XML como medio de transporte.

RDF es un lenguaje de etiquetado que define un modelo de datos para describir recursos, mediante enunciados en forma de tripletas *recurso – propiedad – enunciado*, donde recurso es cualquier cosa que pueda nombrarse mediante una URI (mecanismo propuesto por el W3C, para identificar recursos en la red, cuyo subconjunto más conocido son las URLs); propiedad es una característica o

atributo de un recurso, el cual lleva asociada una URI y puede relacionarse con otras propiedades; por último, el enunciado se encarga de asociar el valor de una propiedad a un recurso. Estos tres elementos son llamados también sujeto, predicado y objeto, respectivamente. Por ejemplo:

Recurso	Propiedad	Enunciado
http://www.paginapersonal.com/recurso.html	Autor	"Antonio Zárate"

```
<?xml version="1.0">
<rdf:RDF>
    <rdf:Description about="http://www.paginapersonal.com/zarate.html">
        <s:autor>Antonio Zárate</s:autor>
    </rdf:Description>
</rdf>
```

Partiendo de RDF, se han desarrollado otros lenguajes para la representación de ontologías:

Ontology Inference Layer (OIL), basado en estándares del W3C (RDF, XML, etc.), representa el conocimiento siguiendo la lógica descriptiva (declaración de axiomas o reglas) y un sistema basado en marcos (taxonomías de clases y atributos) [24].

DARPA Agent Markup Language (DAML) es una propuesta de lenguaje de representación de ontologías, fruto del trabajo conjunto de los grupos de trabajo de OIL y DARPA (que previamente habían desarrollado DAML, otro lenguaje para el uso de ontologías), cuyo objetivo es el de extender la expresividad del RDF pero que presentó cierta complejidad conceptual de uso [6].

OWL (Web Ontology Language) es un trabajo del grupo WebOnt del W3C, el cual trata de un lenguaje derivado de DAML+OIL, a su vez cimentado sobre RDF, que busca solucionar la complejidad conceptual de que adolecía su antecesor. OWL permite definir clases (conceptos), propiedades y ejemplares; además contiene una serie de predicados predefinidos que ayudan en la definición de ontologías. OWL es el lenguaje estándar propuesto por el W3C en 2004 para el diseño de ontologías y se divide en tres sublenguajes [37]:

- OWL Full. Es la unión de la sintaxis de OWL y RDF sin restricciones, lo que permite una alta expresividad, a costo de una mayor dificultad para trabajar con las ontologías, en cuanto a la conceptualización, decisiones y consultas que se le puedan hacer.

- OWL DL (Description Logics). En este sublenguaje se limita la expresividad en favor de las decisiones. La gran mayoría de las herramientas que trabajan sobre ontologías soportan este sublenguaje, siendo un gran punto a su favor.

- OWL Lite. Es un subconjunto de OWL DL más sencillo de implementar, mediante el manejo de una menor complejidad formal en comparación con los otros sublenguajes mencionados.

CAPÍTULO 3 Metodología de solución

El capítulo aborda un análisis más detallado de la problemática a la que se hizo frente, abarcando las distintas etapas realizadas, la construcción de la ontología, la clasificación de elementos dentro de ella y la navegación a través de la misma.

3.1 Metodología de configuración propuesta

La metodología de configuración propuesta para una ILNBD se basa en la definición de los ejemplares y sus relaciones que le sirvan a la interfaz en su proceso de traducción: partes del habla e identificadores de los elementos de la base de datos. Por ejemplo: si se tiene en la base de datos una tabla llamada “sales”, lo correcto es que se relacione con sustantivos como “ventas”, “pedidos”, etc. y con verbos como “comprar”, “vender”, etc. Para que la interfaz trate de “entender” que entre “sales” y “ventas” hay una relación del identificador de una tabla con respecto a una palabra, y que entre conceptos como “ventas” y “pedidos” hay una relación de sinonimia o equivalencia en su significado, es necesario definir estas relaciones y su significado.

Como se comentó anteriormente, los conceptos y las relaciones deben estar organizados, ya que sin esto, es muy difícil utilizarlos, proponiéndose una ontología como modelo de organización. Dos principios clave de las ontologías son el reuso y la compartición de recursos, por lo cual, es necesario que la organización que tengamos sea lo más genérica posible, para que varios usuarios la puedan compartir, y a su vez, basarse en principios generalmente aceptados, de tal manera que también varios usuarios la puedan comprender y reusar. Esto es muy útil porque el objetivo de las interfaces de lenguaje natural es que sean utilizadas por un gran número de usuarios, que justifique lo costoso de su configuración, operación y mantenimiento.

Para lograr el objetivo de que la ontología fuese lo más genérica posible, se recurrió a la lingüística y a la gramática [32] como guías para establecer las categorías en las que se organizarían conceptos y relaciones. A su vez, nos basamos en la teoría de bases de datos [4], para establecer las categorías o clases de los elementos que conforman la base de datos. No hay que olvidar que la traducción de una consulta en lenguaje natural hacia una base de datos se puede entender como la búsqueda de relaciones que asocien los elementos de la consulta (sustantivos, adjetivos, etcétera) con elementos de la base de datos, tal que permitan expresar la consulta en un lenguaje formal (en nuestro caso SQL).

Como elementos adicionales, se agregaron categorías y relaciones enfocadas a la operación del prototipo de ILNBD, como por ejemplo la categoría *función definida por el usuario*, que permite especificar que la definición de una frase o concepto nuevo equivale a una función o procedimiento de la base de datos. Por ejemplo, la consulta “dame el sueldo total de un empleado”, puede ser equivalente a un procedimiento llamado *calcula_sueldo_empleado.*

Para lograr que la ontología, que sirve como base de conocimiento de la ILNBD, fuese lo más reutilizable posible, se formalizó en el Lenguaje de Ontologías para la Web (OWL). Lo anterior nos garantiza que se pueden usar otras ontologías formalizadas en OWL como base de la ontología o compartirla con otros usuarios a través de la Web, para que reutilicen los conceptos y relaciones en otras aplicaciones o con la ILNBD en un contexto semejante.

3.2 Creación de un dominio

A semejanza de otras metodologías definidas por ILNBDs comerciales [7], [20], la propuesta incluye la definición del dominio de la base de datos con la que va a trabajar la interfaz. Esta definición es recomendable hacerla en varias fases que se describen a continuación:

Analizar la semántica de la base de datos. En caso de no ser el diseñador o administrador de la base de datos, es necesario comprender de qué se trata cada uno de los elementos de la base de datos (tablas, columnas, vistas, etc.) y con qué conceptos del lenguaje natural (sustantivos, verbos, etc.) están relacionados, para lo cual es necesario revisar el esquema de la base de datos, la descripción de tablas y columnas, y la documentación de la base de datos, en caso de que se cuente con ello. Un ejemplo de lo anterior sería encontrar en la documentación de la base de datos *Nortwind* que *Quantity* de la tabla *Sales* se refiere a la cantidad de productos que se venden en un pedido.

Obtener un corpus de consultas de los usuarios potenciales. Recolectar un corpus de consultas de una muestra representativa de nuestros usuarios potenciales (se aconseja que sean por lo menos diez consultas por cada usuario),

Clasificar el corpus en categorías. Hacer un análisis del corpus recolectado para clasificar las consultas a partir de la información implícita y explícita que proporcionan con respecto a la información contenida en el diccionario de datos, para definir la información que es necesario proporcionar a la ILNBD, para que responda las consultas de los usuarios potenciales. Dos ejemplos sencillos serían los siguientes: “¿Cuáles son las ventas del empleado Smith?” y “¿Cuántos empleados hay contratados?”. La primera se clasificaría como una consulta que tiene explícitamente definidas las tablas (*Empleado* y *Ventas*), para lo cual sólo es necesaria la información del diccionario de datos. La segunda también está en la categoría de consultas con tablas explícitamente definidas (empleados). Para ésta se puede aplicar la solución dada anteriormente, y sólo es necesario relacionar la palabra de la consulta “Cuánto” con la función de agregación contar (*count* en SQL), y definir que un empleado contratado es aquél que tiene fecha de contratación (*hiredate)* no nula. Esta actividad les permitiría a los usuarios finales esperar que por lo menos las consultas que hagan, semejantes a las proporcionadas anteriormente para el corpus, sean contestadas de forma correcta por la ILNBD.

Definir los conceptos útiles para responder las consultas. No todos los elementos de una consulta tendrían que ser agregados a la ontología, ya que varios de ellos no varían con respecto a un contexto, por lo cual se recomienda descartar artículos, conjunciones, preposiciones (categoría otros) y partes de la consulta relacionadas con la condición de selección como elementos a configurar

para una base de datos; igualmente, se recomienda descartar palabras que no contribuyan hacia el significado de la consulta; por ejemplo, la palabra *muéstrame* en la consulta “Muéstrame los libros que han sido más leídos”. El resto de las partes del habla a configurar (sustantivos, verbos, adjetivos y adverbios) deben ser lemantizados, ya que no tendría sentido configurar en la ontología variantes de una palabra, por ejemplo, se configuraría la palabra *leer*, no las palabras *leídos*, *leídas*, *leído*, *leída*, etc.

Identificar las relaciones y conceptos en los ejemplares de la ontología prellenada. Revisar la información léxica y sintáctica que viene inmersa en la ILNBD y revisar la información semántica de la ontología predefinida. Ésta consiste en la información del diccionario, un conjunto de conceptos, sus sinónimos, y un conjunto de relaciones adicionales (hiperonimia, meronimia, causa-efecto, etcétera) entre dichos conceptos y el tratamiento para manejar cada una de dichas relaciones. Dentro de la ILNBD se puede incluir la implementación de la estrategia para tratar cada relación, pero también se permite que el usuario defina sus propias relaciones y las enlace a programas implementados por él, para agregarle funcionalidad a la ILNBD, donde estos conceptos y sus relaciones constituirían una extensión a la ontología predefinida. Para revisar la ontología predefinida y configurar la ILNBD, se puede utilizar el editor de ontologías Protégé [35] o cualquier editor que soporte a OWL.

Conectar los elementos de la consulta con los conceptos y relaciones que explican la semántica de la base de datos. Después de cargar la ontología predefinida, revisar el análisis del corpus de consultas y comparar la información que se necesita para poder contestar las consultas de cada categoría, con la información contenida en la ontología. Como resultado de esta comparación, puede concluirse, en el mejor de los casos, que la ontología predefinida ya contiene la información para poder contestar algunas consultas, o que para otras, sólo es necesario agregar algunos ejemplares que se pueden asociar por alguna de las relaciones predefinidas (en la sección 3.4 se explica más a detalle).

Agregar las palabras faltantes como ejemplares de la clase *Palabras* y relacionarlas a synsets. Este punto se detalla en la sección 3.6 “Ejemplares”.

Si una palabra no tiene synset con el cual asociarse, agregar un synset y enlazarlo con sus elementos correspondientes. Este punto se detalla en la sección 3.6 “Ejemplares”.

Agregar y asociar las relaciones faltantes al programa que implementará su semántica. Si como resultado del análisis anterior se concluye que las clases y relaciones predefinidas en la ontología no son suficientes para contestar los tipos de consultas obtenidas del corpus, es necesario determinar qué relaciones y clases es necesario agregar a la ontología, como parte de la funcionalidad de la

ILNBD. En este punto es recomendable definir nuevas relaciones, que serán interpretadas en base a las *funciones definidas por el usuario.* Este proceso se describe más a detalle en la sección 3.5 “Definiendo una nueva relación”.

3.3 Clases (categorías), conceptos (synsets) y palabras

Como se mencionó anteriormente, la ontología implica una serie de categorías o clases, para poder organizar los conceptos que definen el contexto de la base de datos con la que va a trabajar la ILNBD. En la figura 3-1 se pueden ver las clases principales de la ontología predefinida, siendo éstas categorías abstractas, no los nombres de los elementos de la base de datos ni las palabras usadas en las consultas, ni los identificadores de los synsets (concepto tomado de Wordnet [21]) o los nombres de funciones utilizables, ya que en los conceptos se consideran elementos concretos usados como información que necesita la ILNBD y que deben agregarse a la ontología como ejemplares. Una analogía de lo anterior es la definición de las tablas de una base de datos (las clases) y la introducción de los datos (los ejemplares). Como la gran mayoría de todas las ontologías, las clases están organizadas en base a la relación *es un* o subclase. A continuación se explica una definición de las clases principales de la ontología y sus principales subclases:

ElementosBD. Aquí se definen las categorías en las que se clasifican los principales elementos que conforman una base de datos, según el modelo relacional (el más utilizado). Algunas subcategorías fueron omitidas, como índices o disparadores, debido a que un usuario difícilmente va a consultar estos elementos (figura 3-1).

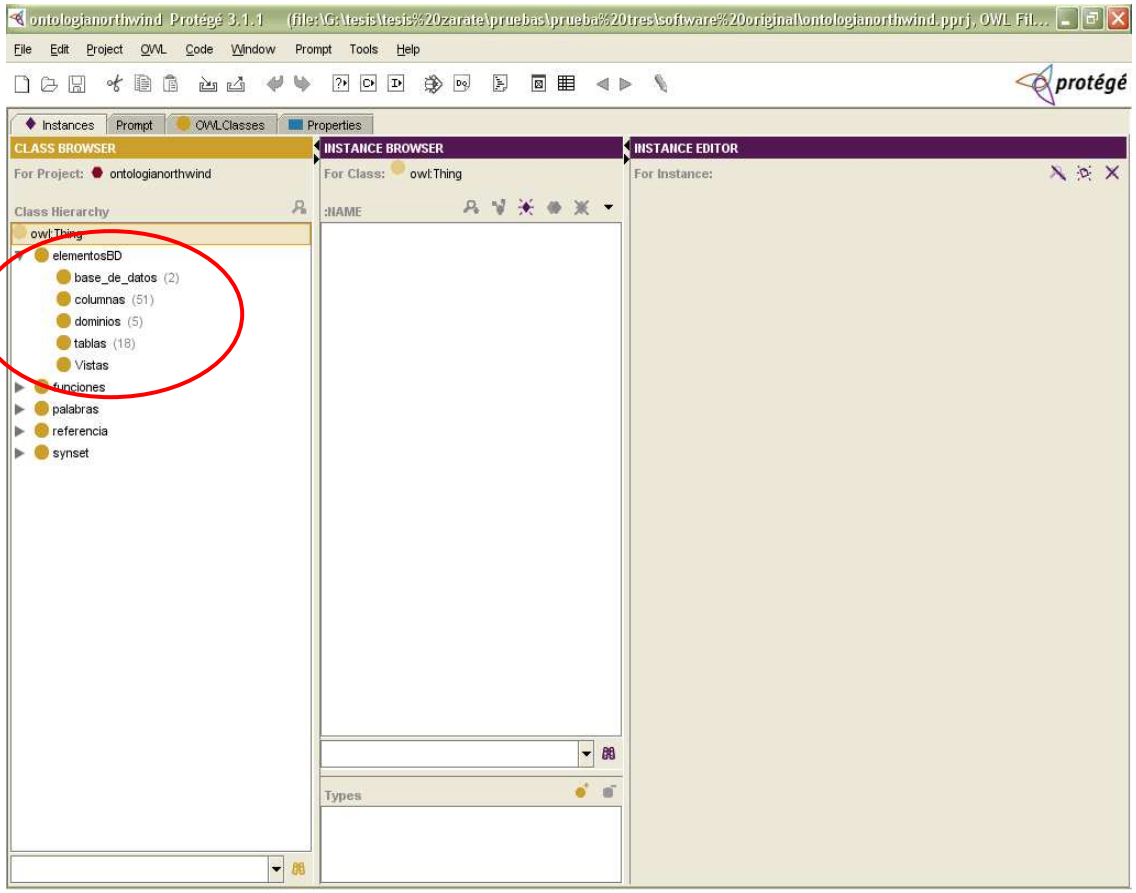

Figura 3.1 Subclases de la categoría *Elementos de la base de datos*.

Funciones. Se dividen en tres subcategorías: las *funciones de agregación* que proporciona el lenguaje SQL, las *funciones definidas por el usuario* y las *relaciones definidas por el usuario*. Ejemplares típicos de la primera subcategoría son nombres de funciones como AVG, MAX, etc.

La clase *funciones definidas por el usuario* permite definir nombres de programas definidos por el usuario que se relacionarán con ejemplares de alguna subclase de *synset*. Un ejemplo aclarativo se muestra en la configuración de la consulta “Dame la antigüedad del empleado Lloyd”. El problema en la consulta es que no existe un campo en donde se almacene explícitamente la antigüedad del empleado, pero sí un procedimiento de la base de datos (stored procedure) llamado *calc_antiguedad*, que calcula dicha antigüedad a partir del campo llamado *fecha de contratación* (en inglés *hiredate*). El ingeniero de conocimiento definiría como ejemplar de la clase *funciones definidas por el usuario* a *calc_antiguedad* y después lo relacionaría con el synset sustantivo *experiencia1* (el cual está relacionado a su vez con los sustantivos *experiencia, antigüedad, decrepitud, vejez, ansiedad* y *arcaísmo*).

Realizar el proceso de la manera anterior ahorra trabajo, ya que si la consulta anterior se cambiará por “Cuál es la experiencia del empleado Lloyd”, no habría necesidad de hacerle cambios a la configuración de la interfaz.

La subcategoría *relaciones definidas por el usuario* sirve para definir etiquetas que funcionan como un puente entre una nueva relación definida por el configurador con un programa externo a la ILNBD que implementa el manejo de la nueva relación (figura 3-2). Un ejemplo sería si el usuario definiera la relación *comprar*, tomando como dominio ejemplares de la relación *comprador* y cuyo rango sería *producto*. Esta nueva relación se podría interpretar vía un programa llamado *interpreta_comprar.class* (el lenguaje y arquitectura del módulo que implementara la traducción no está definido, aquí se define el caso de que este módulo estuviera escrito en el lenguaje Java).

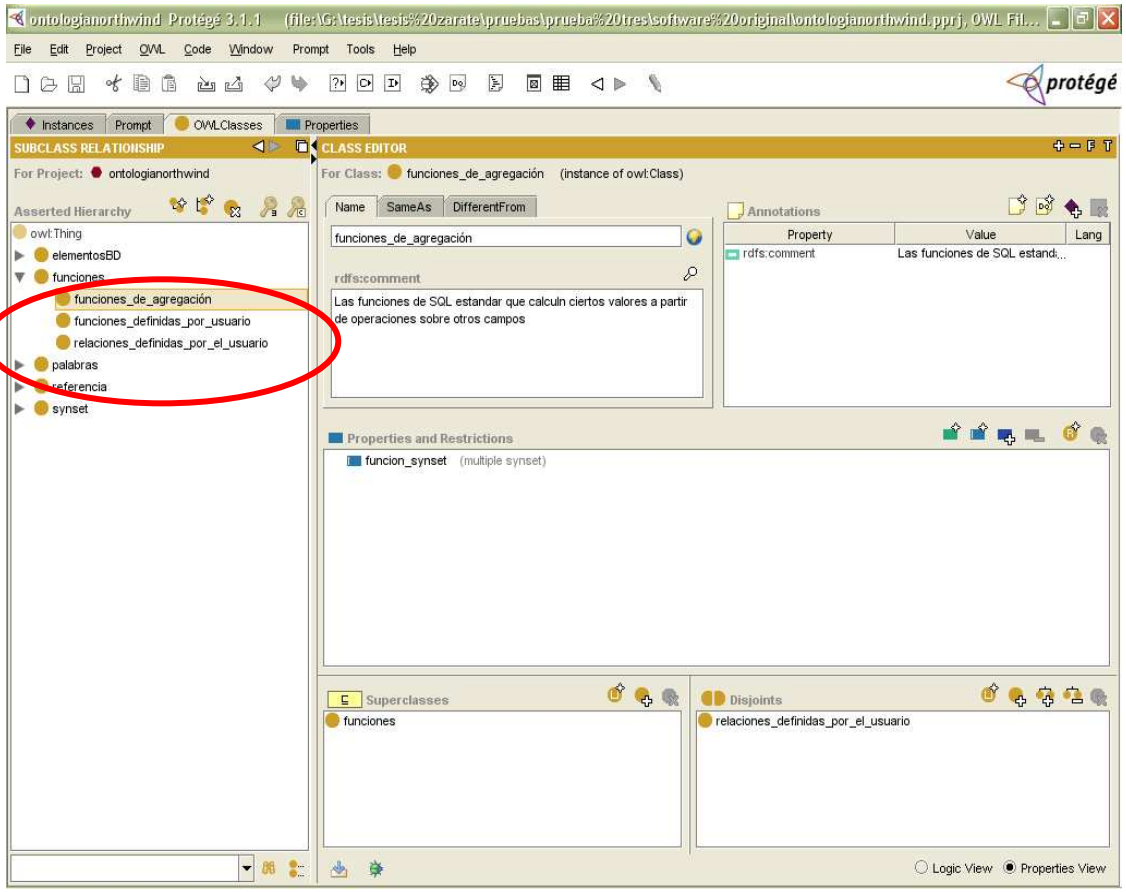

Figura 3.2 Subclases de la categoría *Funciones*.

Palabras. Categoría que representa gráficamente la palabra hablada [27]. Dado que “palabra” normalmente se refiere a la pronunciación y a su concepto asociado, las discusiones de esta asociación léxica son vulnerables a una confusión terminológica. Con el ánimo de reducir ambigüedad, aquí se usará significante

(word form) para referirse a la pronunciación física o la escritura, y significado (word meaning) se usará para referirse al concepto lexicalizado que un significante puede usar para expresar algo. Las subcategorías definidas se basan en lo que en lingüística se conocen como partes del habla (POS por sus siglas en inglés), y en el caso de la subcategoría *otros*, sus subcategorías se basan en la gramática del idioma español (figura 3-3).

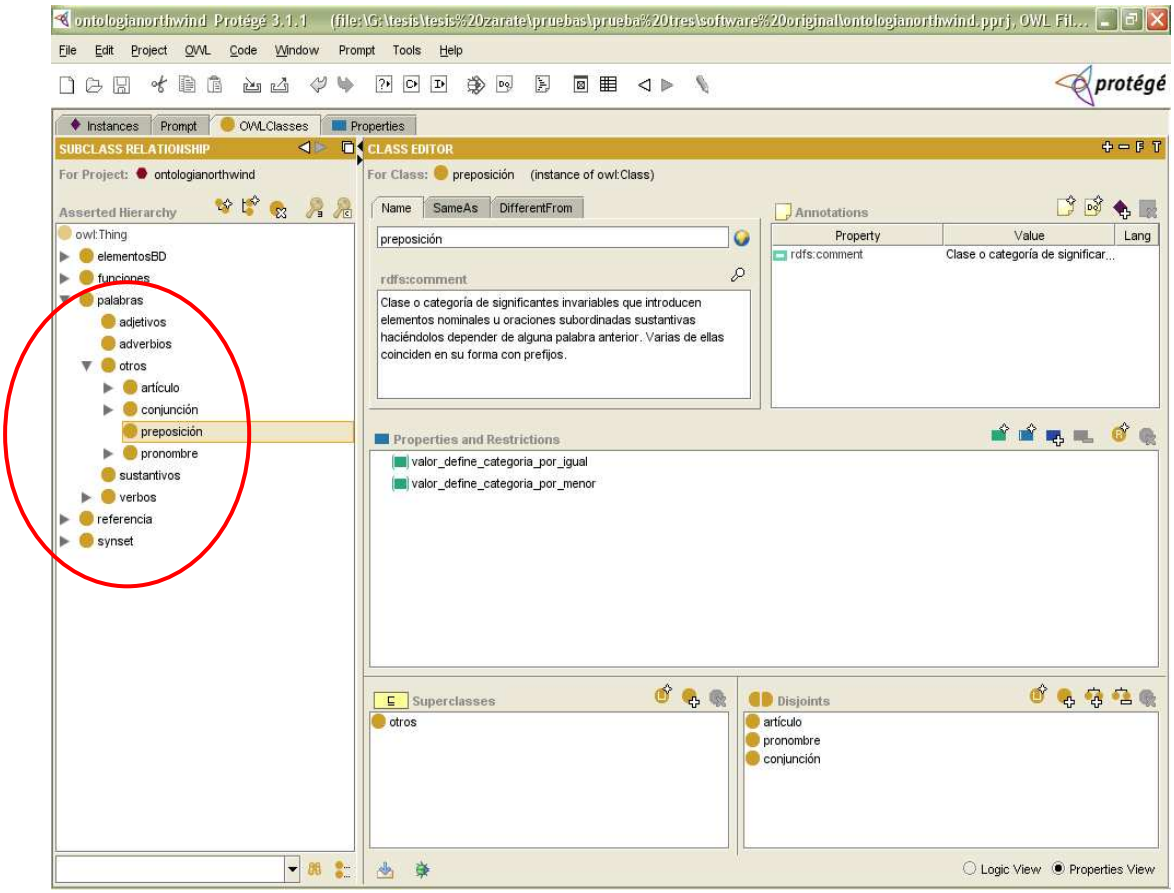

Figura 3-3 Subclases de la categoría *Palabras*.

Synsets. Un synset (contracción de las palabras en inglés *synonym set* o "conjunto de sinónimos" [21]) representa un concepto o significado a través de los sinónimos, es decir, los significantes que "contiene". Los synsets ayudan a solucionar el problema generado por la característica del lenguaje de que un concepto puede ser expresado por diversos significantes equivalentes en significado. Un ejemplo de este problema es el siguiente: al expresar que el significante *maestro* está relacionado por *enseñar* con el significante *alumno*, también se tiene que tomar en cuenta que si *maestro* está relacionado por sinonimia con el significante *profesor*, el significante *profesor* está relacionado por *enseñar* con el significante *alumno*, y que para complicar la situación, también se tiene que tomar en cuenta que *alumno* está relacionado por sinonimia con el

significante *estudiante*, lo cual conduciría a que este último también estuviera relacionado por *enseñar* con *profesor* y *maestro*, en resumen, una relación muchos a muchos.

Una solución más sencilla sería enlazar por la relación de sinonimia a *profesor, maestro, catedrático* y el resto de sus sinónimos con una etiqueta (uno de los significantes con un número) como *profesor*1, y hacer lo mismo con *estudiante*, *alumno,* y el resto de sus sinónimos con una etiqueta *estudiante1* (esta etiqueta identifica a un synset) y después enlazar *profesor*1 con *estudiante1*, por medio de la relación *enseñar* una sola en vez, en lugar de relacionar directamente los significantes muchas veces.

Las subcategorías de synset son casi las mismas que las de palabras, excepto la subcategoría “otros”, debido a que estas partes del habla carecen en su mayoría de sinónimos (figura 3-4).

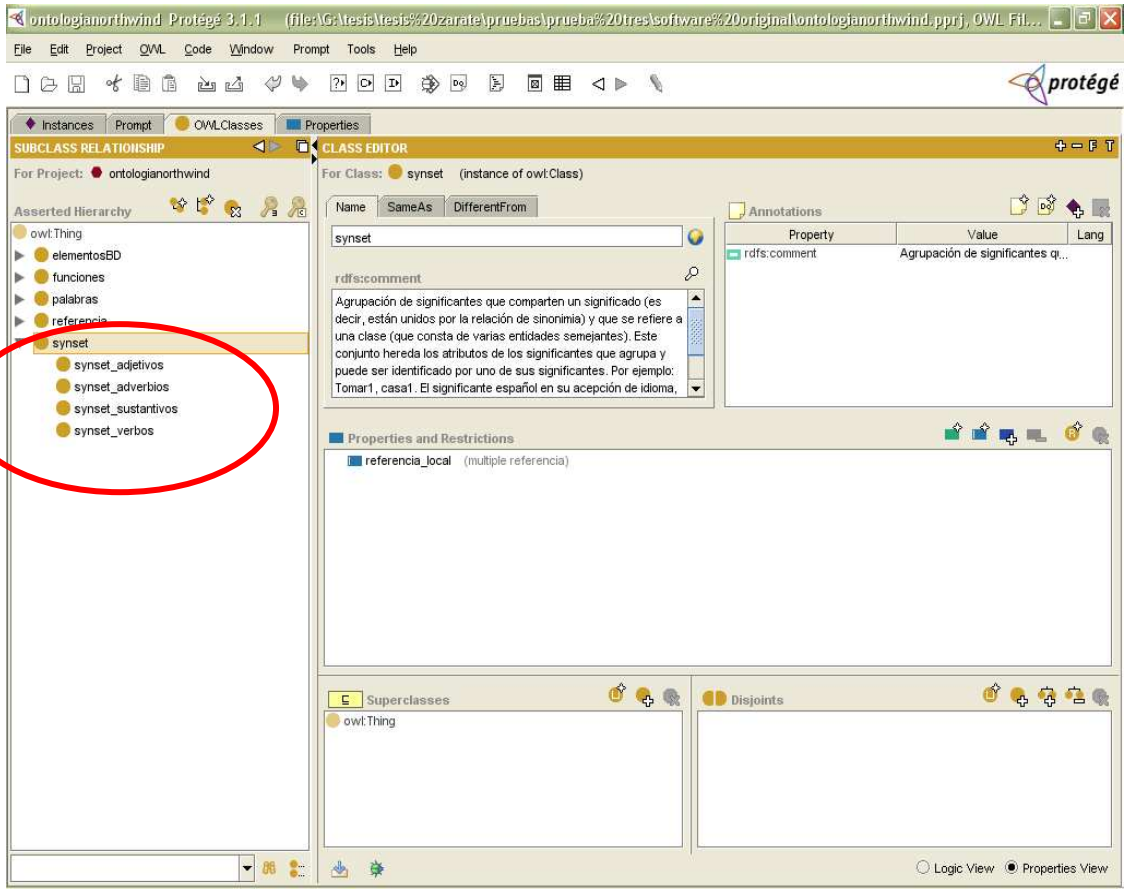

Figura 3-4 Subclases de la categoría *Synset*.

3.4 Relaciones (propiedades)

Son las conexiones o enlaces entre los ejemplares que definen la ontología y, para este trabajo, el contexto de la base de datos que va a ser consultada a través de la ILNBD. Las relaciones pueden verse como enlaces entre un dominio (una clase que define la relación) y un rango (la clase sobre la que se define la relación) y son caracterizadas por propiedades como transitividad, reflexividad, conmutatividad y otras como la herencia, que permite generar subrelaciones. Las principales relaciones definidas en la ontología predefinida son las siguientes:

Relación léxica. Concepto tomado de la lingüística, y definido como un modelo culturalmente reconocido de asociación que existe entre las unidades léxicas en un idioma [32]. La gran mayoría de las relaciones de la ontología predefinida se encuentran aquí (figura 3-5).

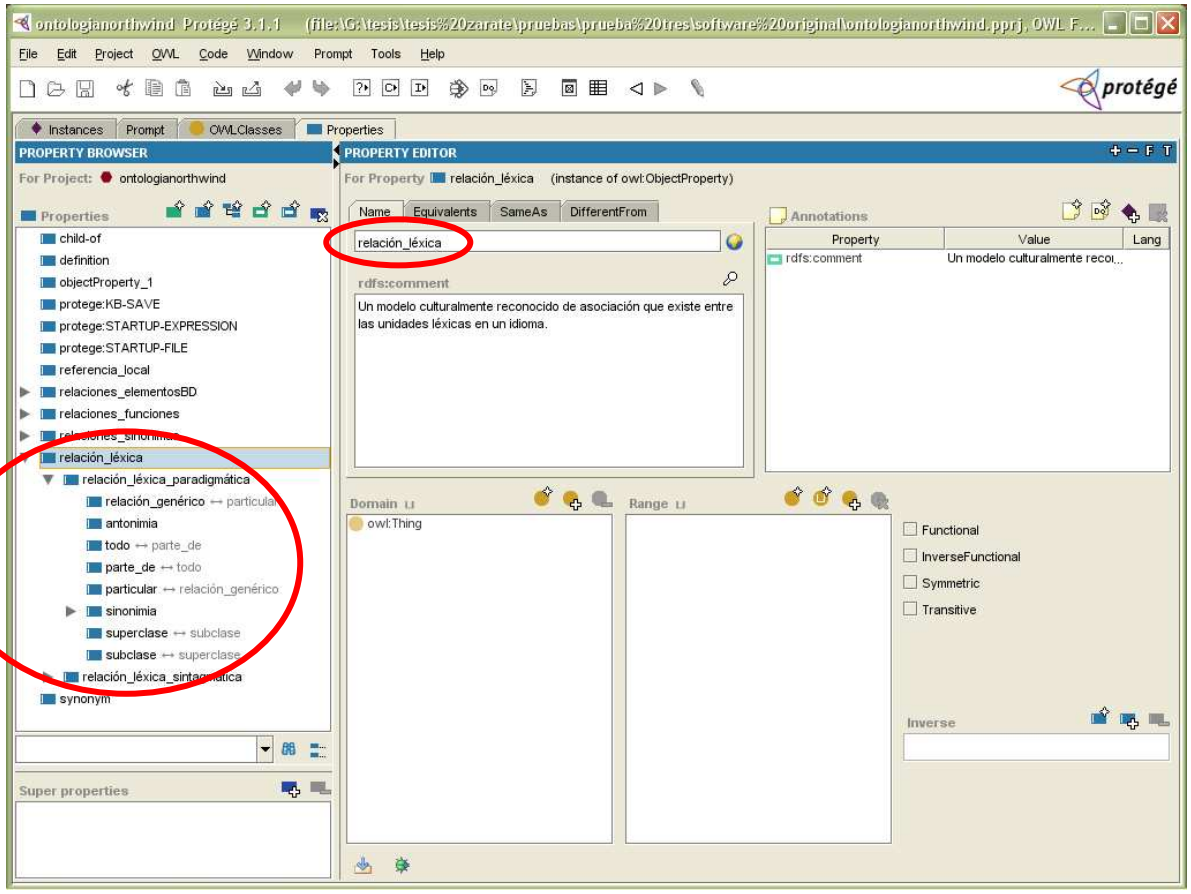

Figura 3-5 Relaciones léxicas.

Relaciones sinónimas. Subrelaciones que son sinónimas de subrelaciones léxico-paradigmáticas a través de la relación de OWL *Equivalents* (figura 3-6).

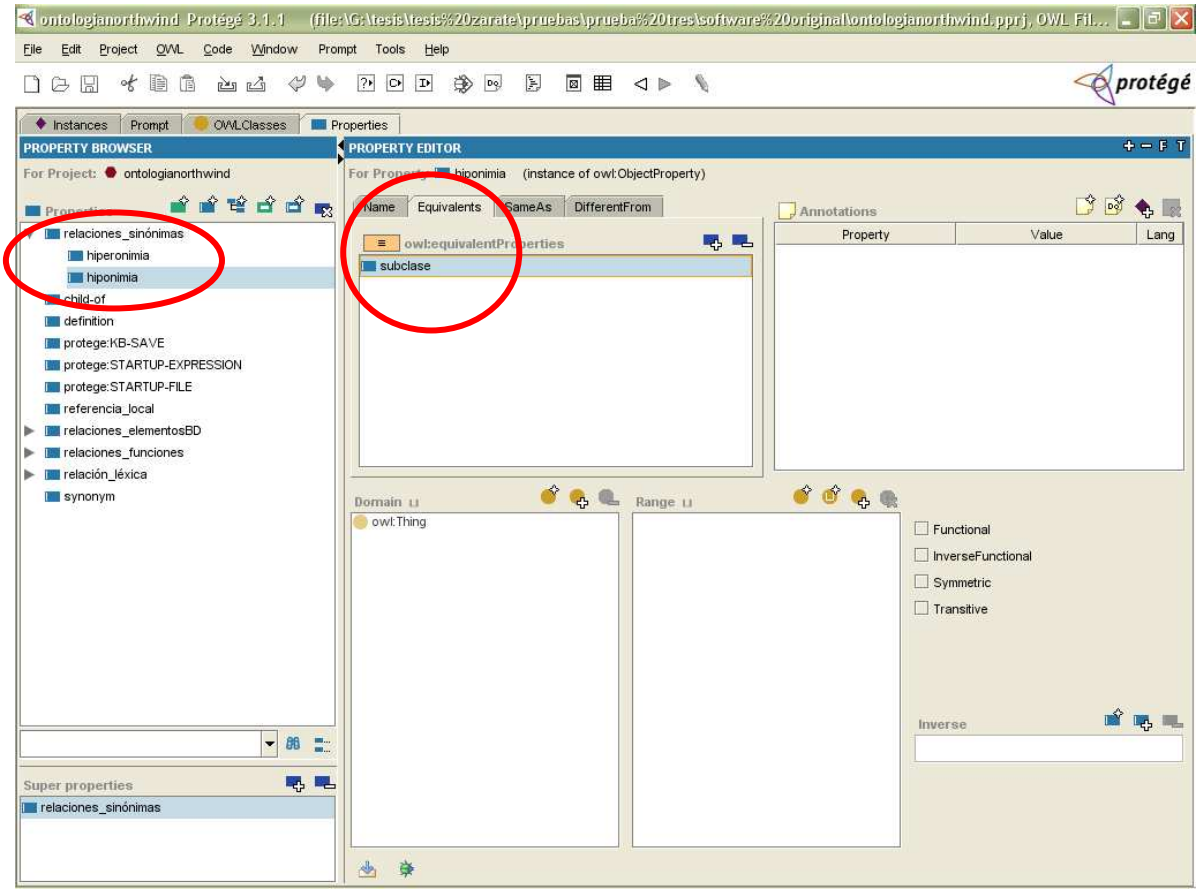

Figura 3-6 Relaciones sinónimas.

Relaciones_elementosBD. Abarca las subrelaciones del modelo relacional (llave primaria, foránea, etcétera) y las relaciones que unen los elementos de la base de datos con synsets, y por transitividad se establece un nexo hacia las *palabras*, las *funciones definidas por el usuario* y *las funciones de agregación* (figura 3-7).

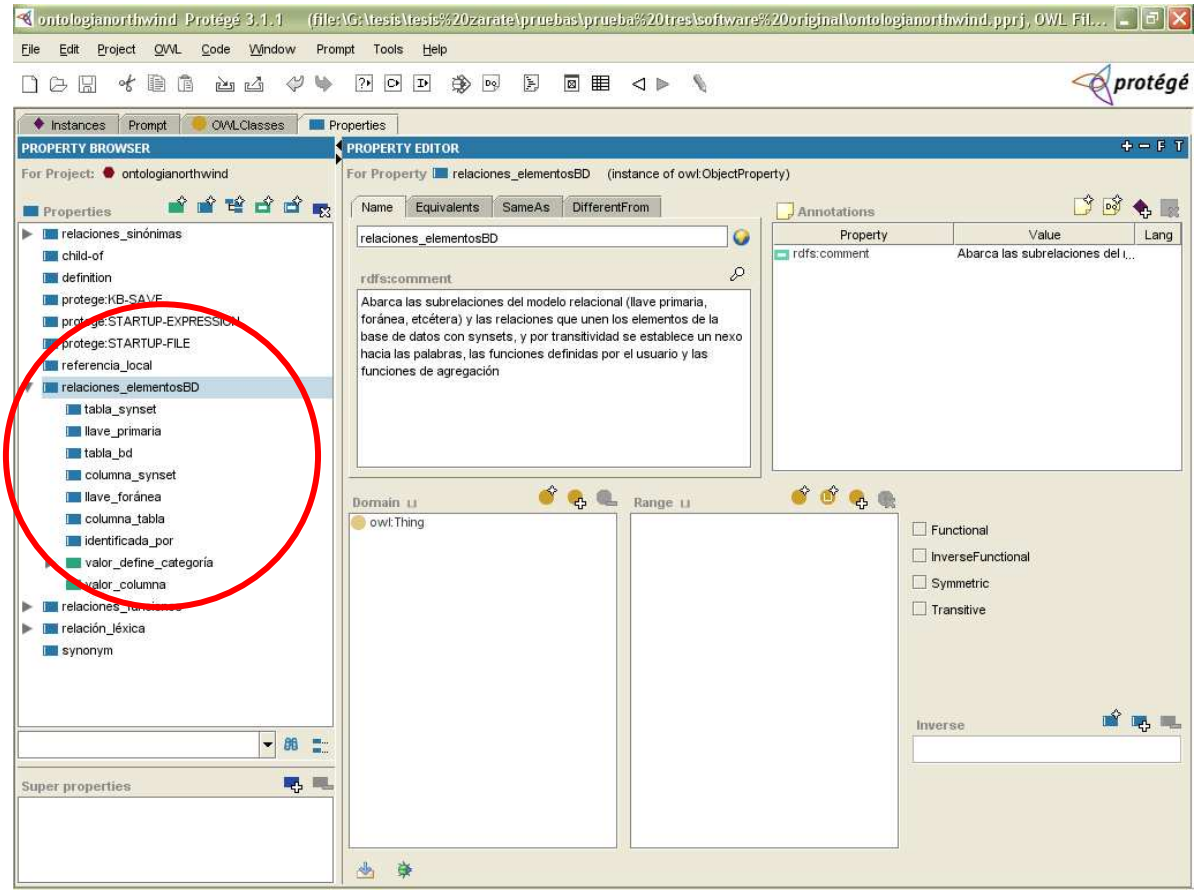

Figura 3-7 Relaciones con elementos de la base de datos.

Relaciones_funciones. Enlaza ejemplares de *funciones definidas por el usuario* con *synsets*, y con nombres de programas (incluyendo su ruta absoluta). Por transitividad con las subrelaciones de *relaciones elementosBD*, los synsets permiten enlazar estas funciones con elementos de la base de datos (figura 3-8).

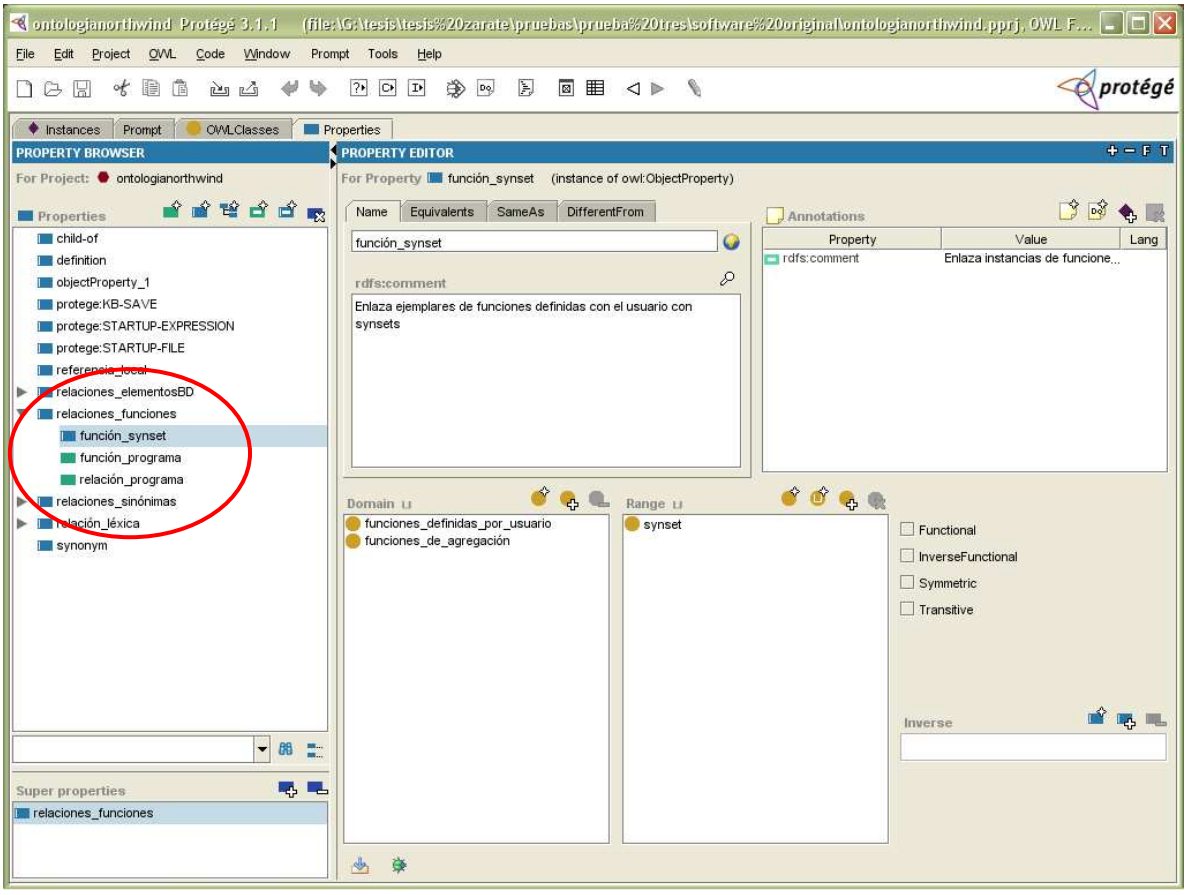

Figura 3-8 Relaciones que permiten utilizar funciones.

3.5 Definición de una nueva relación

La ontología fue diseñada para cubrir las clases y relaciones básicas que implicaría la configuración de una ILNBD, pero también se permite, a semejanza de English Query y ELF, definir relaciones no incluidas en la preconfiguración. La ontología permite agregar a las relaciones definidas por el usuario más funcionalidad que otras ILNBDs, lo cual está restringido únicamente por la expresividad del lenguaje OWL. Este lenguaje permite definir dos tipos de relaciones o propiedades: la propiedad *objeto* (ObjectProperty), en donde su dominio y rango son clases, y la propiedad *datos* (DataProperty), en donde su dominio es una clase y su rango es un tipo definido en el lenguaje XML (figura 3-9).

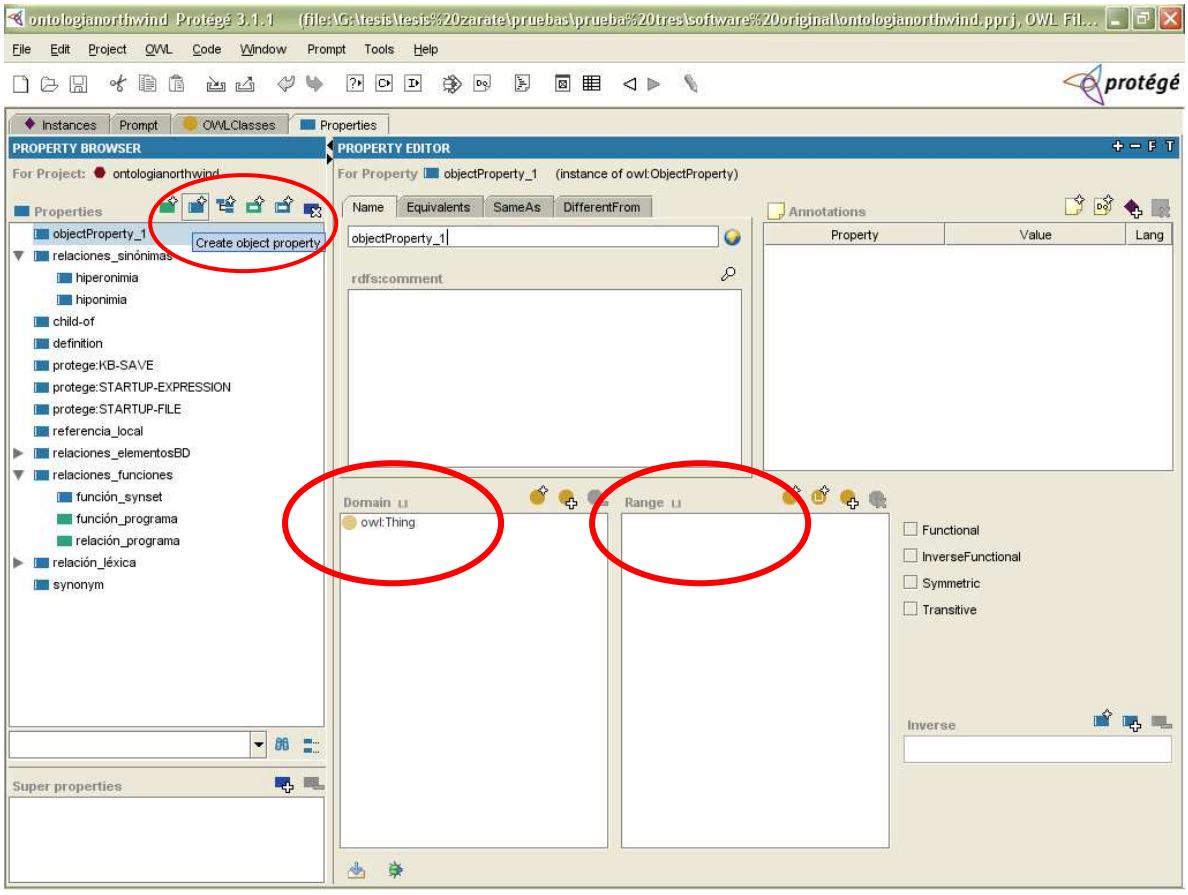

Figura 3-9 Creación de una relación (propiedad).

Es deseable que las relaciones especificadas por el personalizador de la ontología sean a su vez subrelaciones de las relaciones predefinidas, ya que una subrelación hereda las características, dominios y rangos de sus superrelaciones. Esta nueva relación creada por el personalizador no puede ser procesada por la ILNBD y es necesario que se relacione con un programa externo por medio de la relación *SameAs,* que permite enlazar la nueva relación con un ejemplar de la clase *Relaciones_definidas_por_el_usuario.* De esta manera es posible construir el traductor a SQL de una manera modular.

3.6 Ejemplares

Como se mencionó en el apartado 3.3 “Clases (categorías), conceptos (synsets) y palabras”, hasta este momento, sólo se han definido clases y relaciones, no la información que se necesita para que funcione la ILNBD, lo cual es semejante a la definición del esquema de una base de datos, pero sin la captura de los datos. Un ejemplo de la definición de un ejemplar se muestra en la figura 3-10.

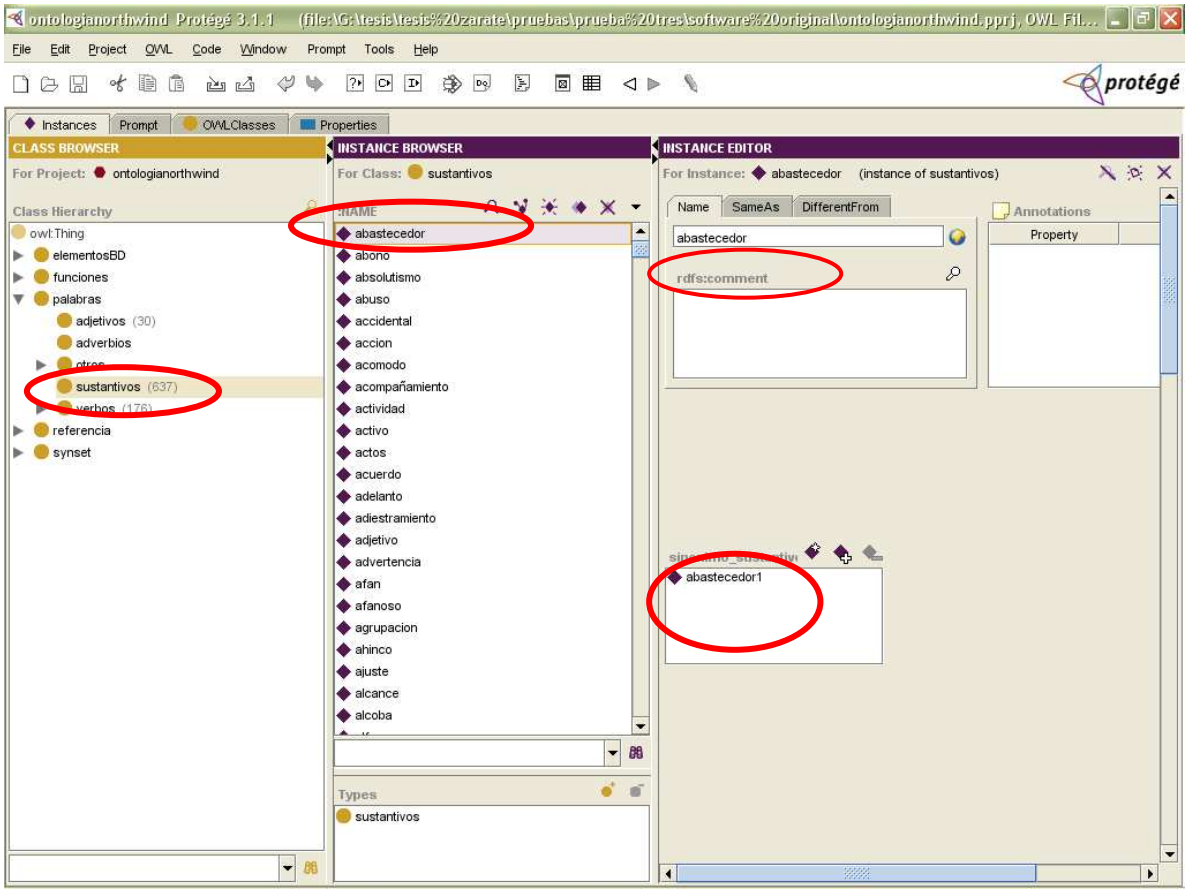

Figura 3-10 Definición de un ejemplar.

Las relaciones de un ejemplar se definen a partir del concepto dominio-relación-rango (explicado anteriormente en la sección 3.4). Cuando se define un ejemplar, se toma éste como el valor constante para todas las relaciones en cuya definición aparece como dominio, y sólo es necesario definirle el ejemplar o ejemplares tomados de la clase definida en el rango de la relación. Tomando el ejemplo de la figura 3-10, *Address* es un ejemplar de la clase *columnas,* que a su vez es una subclase de la clase *ElementosBD*. *Address* es el dominio de la relación *columna_synset* cuyo rango es el synset *destinatario1*, (una consulta semejante a la de la figura 3-9 nos diría cuáles son las palabras relacionadas con *destinatario1*) y además *Address* es el dominio de la relación *columna_tabla* cuyo rango son los ejemplares de la clase tablas: *Employees, Customers,* y *Suppliers* (lo anterior significaría que *Address* es una columna de las tablas *Employees, Customers* y *Suppliers*).

No es necesario que un ejemplar cumpla como dominio de todas las relaciones asociadas; por ejemplo, puede darse el caso de que una tabla no tenga llave foránea, o de que un sustantivo como *abastecedor*, no tenga antónimo (figura 3-11).

La ventaja de reusar parte de una configuración de una interfaz en lenguaje natural por medio de ontologías se puede apreciar en el siguiente ejemplo: Se tiene una base de datos de una compañía (*Northwind*), donde se tiene una tabla llamada *employee*, y se establece una relación con el synset *empleado1;* éste a su vez se relaciona con el synset *jefe1* por la relación *subordinado* (subrelación de la relación léxica sintagmática). Si se configurara la base de datos *Pubs* (publicaciones), en donde se tiene una tabla *employees* y se desea que tuviera las mismas relaciones que *employee*, solamente se tendría que relacionar con el synset *empleado1*, y las relaciones entre los synsets se podrían reusar prácticamente sin ninguna modificación.

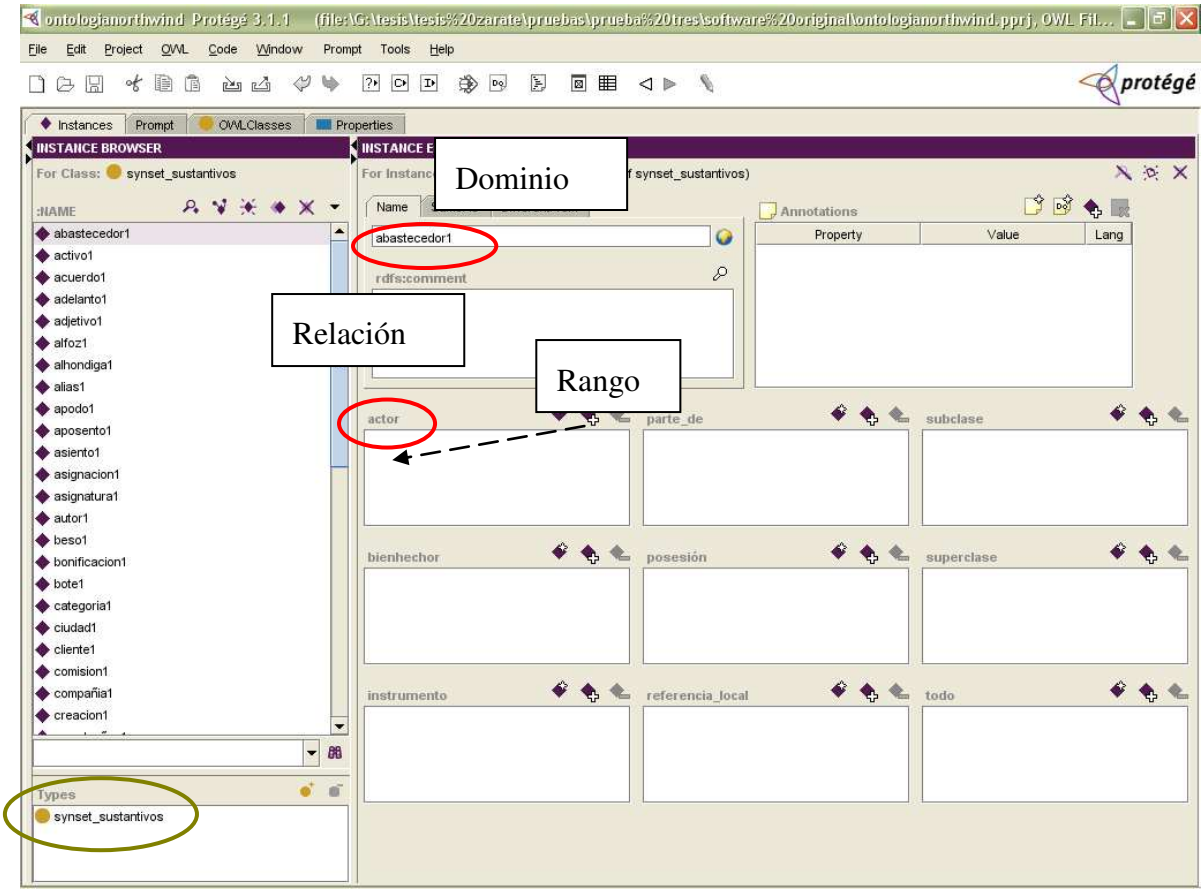

Figura 3.11 Alta de un ejemplar.

CAPÍTULO 4
Validación de la metodología propuesta

En este capítulo se describen las pruebas realizadas para validar la hipótesis propuesta.

4.1 Descripción de la evaluación

Las pruebas no intentaron validar todos los componentes de una ILNBD, ni tampoco validar las respuestas que proporciona, ya que existen muchos factores involucrados: la completez de la base de conocimiento, capacidades de los analizadores sintáctico y semántico, y el tipo de consultas del corpus de prueba (definidas en[12]).

El plan consta de cuatro pruebas de comparación entre la manera en que se configuran English Query (EQ) contra la propuesta de usar ontologías para la configuración de una ILNBD, utilizando el editor de ontologías Protégé [35]. En cada una de estas pruebas se hicieron evaluaciones cruzadas, es decir, un equipo evaluó la propuesta usando Protégé, el otro, EQ; y después se invirtieron los papeles. Se hizo de esta manera debido a lo pequeño de los equipos de prueba y para evitar los posibles sesgos en la conformación de los grupos de prueba.

Entre la primera y la segunda prueba, se hizo una pequeña prueba de afinación del diseño de la ontología, con cinco estudiantes, con el fin de estudiar mejor el proceso de configuración propuesto y en ella se hicieron estudios más detallados (anexo C). En la prueba No. 4 se hicieron tres evaluaciones con alumnos que iniciaban sus estudios de maestría, y que en su mayoría eran recién egresados de licenciatura.

4.2 Grupos de prueba

Los participantes de las pruebas fueron estudiantes de maestría, a los cuales se les dio sólo una plática para explicarles la prueba (no se les capacitó propiamente para evitar posibles sesgos del instructor). Ninguno de los participantes tenía experiencia previa en EQ o Protégé, ni tampoco habían oído hablar de ontologías. Los sujetos de la prueba No. 1 y No. 3 eran de una universidad sin un proceso de admisión riguroso, y por el contrario, los sujetos de las pruebas No. 2 y No. 4 eran de CENIDET, quienes son seleccionados cuidadosamente. Otros datos de cada grupo de prueba se muestran a continuación en la tabla No. 4-1:

Tabla 4-1 Estadísticas de los grupos de prueba.

	Prueba 1	Prueba 2	Prueba 3	Prueba 4
Origen	Universidad privada	CENIDET	Universidad privada	CENIDET
Duración de la prueba	2 hrs.	Mes y medio	Un mes	Semana y media

Corpus de consultas a configurar (dificultad baja/media/alta)	20 (3/9/8)	7 (2/3/2)	8 (3/3/2)	a)10(4/3/3) b)10(1/4/5) c)10(4/3/3)
Documentación disponible	Manual de EQ y un documento describiendo la propuesta y el uso de Protégé.	Manual de EQ, un documento describiendo la propuesta y un documento describiendo algunos ejemplos (sólo texto).	Manual de EQ, un documento describiendo la propuesta y un documento describiendo algunos ejemplos con imágenes.	Igual que la Prueba 3
Número de preguntas del cuestionario de evaluación	7	12	12	12
Promedio de edad	24	24.17	25.1	22.67
Promedio de calificaciones	n/d	92.67	85	93.33
Promedio de años de experiencia	3.35	0.67	2.1	0.5
Número de participantes	17	18	10	6
Sexo	5 M,12 H	9 M, 9 H	2 M, 8 H	4 M, 2 H

Hay muchos otros datos recopilados de los sujetos de prueba como son los cursos tomados, una autoevaluación de sus capacidades de diseño, etcétera, que se muestran en el anexo D *Ejemplo de un cuestionario de obtención de perfil*.

4.3 Tarea a realizar

Se les pidió a los sujetos de prueba que realizaran la configuración de una ILNBD usando Protégé y English Query, para una serie de consultas: veinte en la prueba No. 1, siete para la prueba No. 2, ocho para la No. 3 y diez para cada una las evaluaciones de la prueba No. 4. La dificultad de configuración para cada consulta

varía de baja, media y alta, de acuerdo a qué tanto hay que agregarle a la configuración para que la ILNBD la pueda contestar.

Varias ILNBDs definen sus propios corpus de pruebas [12] y [20], pero se decidió utilizar consultas del corpus de ELF [7] por ser el más usado, tomando como criterio de que la mitad fuera contestable por la configuración predefinida de EQ y la otra mitad no. Un detalle interesante al comparar el corpus de ELF con uno creado por un grupo de prueba, y otro tomado de otro experimento [12], fue que, a pesar de que los tres se referían a la misma base de datos (Northwind), los tipos de consultas que se hacían eran muy diferentes, ya que el primero tiene en su mayoría consultas complejas, el segundo presentó consultas de poca dificultad y el tercero consultas de diversa dificultad. Por último se recopiló un cuarto corpus con consultas que hacen usuarios reales hacia sus bases de datos de operación y la clasificación de las consultas fue diferente en los tipos de consultas realizadas a la de los tres corpus anteriores.

4.4 Cuestionario de evaluación

Las preguntas del cuestionario de la prueba No. 1 se formularon en base a términos de un diccionario relacionados con el concepto *facilidad* [27]. Debido a lo limitado de este experimento, se optó por usar otras métricas más útiles basadas en otro trabajo [11], en donde se evaluaba la mejora de usar una técnica de diseño de bases de datos distribuidas contra otra técnica similar.

A partir de la segunda prueba, las preguntas del cuestionario se agruparon de acuerdo dos categorías principales: preguntas intrínsecas que evalúan la propuesta de usar ontologías como una base de conocimiento de una ILNBD, y preguntas extrínsecas que nos sirven para explicar el porqué se evaluó de esa manera.

Las métricas utilizadas para medir las preguntas del cuestionario se basan en la escala de Likert (uno al siete). Los valores mostrados como resultados de la evaluación de las pruebas (toda la sección 4.5) son valores promedio de las calificaciones de las preguntas del cuestionario de evaluación, normalizadas en una escala de uno al cien. Dos métricas que se usaron en el experimento anterior [11] y no se utilizaron en éste, son el tiempo y la calidad de la configuración obtenida, por las siguientes razones: el tiempo no fue posible medirlo, porque no se pudo reunir a los sujetos de prueba para verificar el tiempo invertido en la tarea de configuración (excepto en la prueba de afinación y en la primera prueba). No se revisó la calidad ni validez de la configuración hecha a partir de la ontología, porque no se contó con un grupo de expertos en el diseño de ontologías (como en el experimento [11]), y además no se pudieron probar las configuraciones producto de la prueba con el analizador semántico disponible [12], porque éste sólo explota la relación de sinonimia y no se quiso restringir que se pudiesen expresar otras relaciones en la definición de la configuración de la base de conocimiento, las cuales no serían interpretadas por el analizador semántico.

4.5 Resultados de la evaluación

4.5.1 Prueba No. 1

Objetivo.- Realizar una prueba exploratoria acerca de la propuesta de utilizar ontologías para configurar una interfaz en lenguaje natural hacia bases de datos, contra la manera de configurar la herramienta comercial English Query, un componente de SQL Server.

4.5.1.1 Resumen de la prueba No. 1

Tabla 4-2 Diferencias entre la propuesta y EQ para la prueba No. 1.

Pregunta	Diferencia
1. Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).	-9.72
2. Considera que la configuración de la interfaz es manejable (moverse con cierta soltura después de haber tenido algún impedimento).	1.52
3. Considera que la configuración de la interfaz es fácil de usar (que se puede hacer sin gran esfuerzo).	-0.63
4. Considera que la configuración de la interfaz es inteligible (que puede ser entendida).	0.00
5. Considera que la documentación de la configuración de la interfaz es fácil de entender (tener idea clara de las cosas).	-1.14
6. Considera que la configuración le da una idea de cómo funciona la interfaz (cómo ejecuta las funciones que le son propias).	6.94
7. Considera que la terminología usada para configurar la interfaz es extraña o confusa.	1.39

Nota: un valor positivo en la columna “diferencia” significa que se consideró mejor la propuesta del uso de ontologías, y un valor negativo significa que se consideró mejor a English Query.

Tabla 4-3 Medidas características para la prueba No. 1.

Medida	English Query	Propuesta
Mejor calificación promedio	63.64 (Considera que la documentación de la configuración de la interfaz es fácil de entender)	63.89 (Considera que la configuración le da una idea de cómo funciona la interfaz)
Peor calificación promedio	54.55 (Considera que la configuración de la interfaz es manejable)	52.78 (Considera que la configuración de la interfaz es flexible)
Diferencia más alta favorable	-9.72 (Considera que la configuración de la interfaz es flexible)	6.94 (Considera que la configuración le da una idea de cómo funciona la interfaz)
Diferencia más baja favorable	0 (Considera que la configuración de la interfaz es inteligible)	0 (Considera que la configuración de la interfaz es inteligible)

4.5.1.2 Análisis de la prueba No. 1

- La mayor parte de los resultados de los promedios normalizados están por encima de la mitad.
- La documentación de English Query es mejor.

4.5.1.3 Conclusiones de la prueba No. 1

Conclusiones positivas:

- Los aspectos en que la propuesta sale mejor evaluada es en dar una idea del funcionamiento de la misma.
- Los comentarios verbales y escritos apuntan a que, a pesar de que no se tenía experiencia con este tipo de interfaces, se entendió su propósito y se percibió en general la flexibilidad de la propuesta de uso de ontologías.
- El concepto básico de ontologías fue en principio bien aceptado.
- No hay grandes diferencias entre las calificaciones dadas, aunque son bajas en general, ya que apenas rebasan la mitad de la escala, debido a la poca capacitación que se les dio a los sujetos de prueba.

Conclusiones negativas:

- El ejercicio de diseño de ontologías fue complicado y dio pocos resultados.
- La segunda parte de la prueba (donde cada uno de los equipos evaluaba la otra ILNBD) fue inconclusa, por falta de tiempo y sus resultados no son muy confiables, porque sólo la mitad de los integrantes de cada equipo la terminó.
- La mayor diferencia en contra se da en el aspecto de flexibilidad, una medida poco confiable, por los aspectos mencionados en el punto anterior.
- Mucha confusión con la tarea de configurar este tipo de interfaces.
- El manual del usuario de English Query era mucho mejor que el manual de la propuesta, debido a la experiencia que tiene la compañía Microsoft en el campo.
- Pocos resultados prácticos al evaluar las configuraciones, tanto de English Query como de la propuesta.

4.5.2 Prueba No. 2

Objetivo.- Realizar una prueba para comparar la propuesta de utilizar ontologías para configurar una interfaz en lenguaje natural hacia bases de datos contra la forma en que se configura English Query, utilizando las experiencias aprendidas en una prueba de afinación (descrita en el anexo B) y en la prueba 1.

4.5.2.1 Resumen de la prueba No. 2

Diferencia entre las evaluaciones

En la tabla 4-4 se muestran las diferencias entre cada una de las preguntas que evalúan el proceso de configuración propuesto basado en una ontología contra el proceso para configurar English Query.

Tabla 4-4 Diferencias entre aspectos intrínsecos que evaluán el proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la prueba No. 2.

Pregunta	Propuesta-EQ
1. Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).	16.67
2. Considera que la configuración de la interfaz es inteligible (que puede ser entendida).	20.83
3. Considera que la configuración le da una idea de cómo funciona la interfaz (cómo ejecuta las funciones que le son propias).	26.04
4. Se sintió cómodo al analizar y completar los conceptos de la ontología.	18.75
5. Se sintió cómodo al analizar y completar las relaciones de la ontología.	21.88

En la tabla 4-5 se muestran las diferencias entre cada una de las preguntas que nos describen la evaluación del proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-5 Diferencias entre aspectos extrínsecos que describen la evalución del proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la prueba No. 2.

Pregunta	Propuesta-EQ
1. Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado.	14.58
2. La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz.	18.75
3. Me sentí cómodo con el ambiente de trabajo después de la sesión de entrenamiento.	21.88
4. Fue sencillo aprender a usar el editor Protégé o el ambiente de configuración de English Query (según sea el caso).	23.96
5. Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo.	23.96

6. Considera que la documentación de la configuración de la interfaz es fácil de entender (tener idea clara de las cosas).	21.88
7. Considera que la terminología usada para configurar la interfaz es extraña o confusa.	18.75

Nota: un valor positivo en la columna Propuesta-EQ, significa que se consideró mejor la propuesta del uso de ontologías, y un valor negativo significa que se consideró mejor a English Query.

Tabla 4-6 Medidas características para la prueba No. 2.

Medida	English Query	Propuesta
Mejor calificación promedio	62.50 (Considera que la configuración de la interfaz es flexible)	79.17 (Considera que la configuración le da una idea de cómo funciona la interfaz)
Peor calificación promedio	44.79 (Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo)	65.63 (Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado)
Diferencia más alta favorable	Ninguna	26.04 (Considera que la configuración le da una idea de cómo funciona la interfaz)
Diferencia más baja favorable	Ninguna	14.58 (Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado)

4.5.2.2 Análisis de los resultados de la prueba No. 2

- En esta prueba los promedios de calificaciones de todas las preguntas favorecen a la propuesta.
- Las mejores calificaciones para la propuesta se logran en los aspectos de flexibilidad y comprensión del funcionamiento de la interfaz, siendo esto último congruente con la evaluación No. 1.
- Las calificaciones más bajas fueron en la evaluación del entrenamiento, lo cual concuerda con el hecho de que no se les proporcionó un entrenamiento formal.

4.5.2.3 Conclusiones de la prueba No. 2

Conclusiones positivas:

- El hecho de que con la propuesta se tenga una mejor comprensión del funcionamiento de la interfaz, tanto en la prueba No. 1 como en la prueba No. 2, es muy importante, porque para extender la configuración de una ILNBD, primero hay que entender cómo funciona.
- Los comentarios verbales y escritos apuntan a que, a pesar de que no se tenía experiencia con este tipo de interfaces, se entendió su propósito y se percibió en general la flexibilidad de la propuesta de uso de ontologías, siendo este aspecto el segundo mejor evaluado.
- La documentación para explicar la propuesta fue bien recibida.

Conclusiones negativas:

- En esta prueba se tuvieron problemas al usar la interfaz de EQ, porque las relaciones y toda la configuración está en inglés, un aspecto reflejado en la pregunta “Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo”.
- A pesar de que se solicitó a los sujetos de prueba una descripción del proceso de configuración, tanto para la propuesta como para EQ, se tuvieron pocos resultados para determinar qué tan bien se asimiló el concepto de configuración de una ILNBD.
- Se agregó un tutorial en español de diseño de ontologías, desarrollado por una experta en el campo [23], pero no fue usado por los estudiantes, debido a problemas de motivación.
- Hubo mucha confusión con la tarea de configurar este tipo de interfaces.
- La prueba se demoró demasiado debido a que se atravesó un periodo vacacional.

4.5.3 Prueba No. 3

Objetivo.- Realizar una prueba para comparar la propuesta de utilizar ontologías para configurar una interfaz en lenguaje natural hacia bases de datos contra la forma en que se configura English Query, con el fin de corroborar los resultados obtenidos en la prueba No. 2.

4.5.3.1 Resumen de la prueba No. 3

En la tabla 4-7 se muestran las diferencias entre cada una de las preguntas que evalúan el proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-7 Diferencias entre aspectos intrínsecos que evalúan el proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la prueba No. 3.

Pregunta	Propuesta-EQ
1. Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).	14.58
2. Considera que la configuración de la interfaz es inteligible (que puede ser entendida).	10.42
3. Considera que la configuración le da una idea de cómo funciona la interfaz (cómo ejecuta las funciones que le son propias).	12.50
4. Se sintió cómodo al analizar y completar los conceptos de la ontología.	6.25
5. Se sintió cómodo al analizar y completar las relaciones de la ontología.	8.33

En la tabla 4-8 se muestran las diferencias entre cada una de las preguntas que describen la evaluación del proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-8 Diferencias entre aspectos extrínsecos que describen la evalución del proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la prueba No. 3.

Pregunta	Propuesta-EQ
1. Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado.	0.00
2. La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz.	4.17
3. Me sentí cómodo con el ambiente de trabajo después de	8.33

la sesión de entrenamiento.	
4. Fue sencillo aprender a usar el editor Protégé o el ambiente de configuración de English Query (según sea el caso).	-2.08
5. Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo.	25.00
6. Considera que la documentación de la configuración de la interfaz es fácil de entender (tener idea clara de las cosas).	8.33
7. Considera que la terminología usada para configurar la interfaz es extraña o confusa.	6.25

Nota: un valor positivo en la columna Propuesta-EQ, significa que fue mejor la propuesta del uso de ontologías, y un valor negativo significa que se consideró mejor a English Query.

Tabla 4-9 Medidas características para la prueba No. 3.

Medida	English Query	Propuesta
Mejor calificación promedio	66.67 (Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado)	75.00 (Considera que la configuración de la interfaz es flexible)
Peor calificación promedio	50 (Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo)	58.33 (Fue sencillo aprender a usar el editor Protégé o el ambiente de English Query)
Diferencia más alta favorable	-2.08 (Fue sencillo aprender a usar el ambiente de configuración de English Query)	25.00 (Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo)
Diferencia más baja favorable	0 (Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado)	0 (Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado)

4.5.3.2 Análisis de los resultados de la prueba No. 3

- La mayor parte de los resultados de los promedios normalizados están por encima de los promedios de la prueba No. 2.
- En la mayor parte de las preguntas, la propuesta es mejor que EQ, ya que de las doce preguntas del cuestionario, la propuesta es mejor en diez, sólo es menor en una y en sólo una pregunta los resultados son iguales.
- El único aspecto en el que EQ sale mejor evaluado es en que es más sencillo de aprender a usar.
- En el aspecto en que salen igualmente evaluados EQ y la propuesta es la apreciación del entrenamiento recibido.
- Las mejores y las peores calificaciones son diferentes de las de la prueba No. 2, aunque en general la propuesta continúa bien evaluada en aspectos como lo entendible y flexible que es.

4.5.3.3 Conclusiones de la prueba No. 3

Conclusiones positivas:

- El aumento en el valor de los promedios puede explicarse por el mayor tiempo disponible para practicar con EQ y Protégé (se verificó que por lo menos tres horas le dedicaran a practicar con estas herramientas y a leer la documentación).
- Se tienen problemas de comprensión del inglés con los sujetos de esta prueba (valores de 50 y 75 en la pregunta No. 5 de aspectos intrínsecos, relacionada con los problemas que implica usar una interfaz en un idioma diferente al español).
- Al no haber diferencia en la percepción de lo adecuado del entrenamiento recibido para usar las dos ILNBDs, se puede asegurar que se evita en lo posible el sesgo inherente del aplicador de la prueba, siendo un factor fundamental que en esta prueba se tuvo mayor tiempo de práctica tanto con EQ como con Protégé.
- En general, en los aspectos relacionados con la portabilidad de la interfaz (flexible, entendible, etc.) la propuesta sale mejor evaluada.

Conclusiones negativas:

- Por comentarios verbales, se considera que la configuración de la propuesta es más tediosa, ya que al principio lleva más tiempo que EQ. Esta deficiencia de la propuesta es superada al reusar parte de la configuración al portar la interfaz a un contexto semejante, o al clasificar las

preguntas en patrones y reciclar el procedimiento para solucionarlas, pero lamentablemente esta apreciación no fue posible validarla.

- A los evaluadores se les hizo más claro y sencillo aprender EQ, debido a que se parece mucho a los productos de la misma compañía que varios de ellos utilizan diariamente en sus trabajos (su promedio de años de experiencia laboral es mayor que la de los sujetos de la prueba No. 2).

4.5.4 Prueba No. 4

4.5.4.1 Evaluación No. 1 de la prueba No. 4

Objetivo.- Realizar una prueba con un pequeño grupo de egresados de licenciatura, recién ingresados a la maestría, para conocer el comportamiento al utilizar ontologías para configurar una interfaz en lenguaje natural hacia bases de datos versus el procedimiento para configurar English Query, y evaluar si los conocimientos de un estudiante de licenciatura son suficientes para configurar adecuadamente una ILNBD.

4.5.4.1.1 Resumen de la evaluación No. 1 de la prueba No. 4

En la tabla 4-10 se muestran las diferencias entre cada una de las preguntas que evalúan el proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-10 Diferencias entre aspectos intrínsecos que evalúan el proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la evaluación No. 1 de la prueba No. 4.

Pregunta	Propuesta-EQ
1. Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).	8.33
2. Considera que la configuración de la interfaz es inteligible (que puede ser entendida).	-37.50
3. Considera que la configuración le da una idea de cómo funciona la interfaz (cómo ejecuta las funciones que le son propias).	-25.00
4. Se sintió cómodo al analizar y completar los conceptos de la ontología.	-12.50
5. Se sintió cómodo al analizar y completar las relaciones de la ontología.	-12.50

En la tabla 4-11 se muestran las diferencias entre cada una de las preguntas que describen la evaluación del proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-11 Diferencias entre aspectos extrínsecos que describen la evalución del proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la evaluación No. 1 de la prueba No. 4.

Pregunta	Propuesta-EQ
1. Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado.	-8.33
2. La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz.	-4.17
3. Me sentí cómodo con el ambiente de trabajo después de la sesión de entrenamiento.	-12.50
4. Fue sencillo aprender a usar el editor Protégé o el ambiente de configuración de English Query (según sea el caso).	-12.50
5. Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo.	16.67
6. Considera que la documentación de la configuración de la interfaz es fácil de entender (tener idea clara de las cosas).	-8.33
7. Considera que la terminología usada para configurar la interfaz es extraña o confusa.	-25.00

Nota: un valor positivo en la columna Propuesta-EQ, significa que se consideró mejor la propuesta del uso de ontologías, y un valor negativo significa que se consideró mejor a English Query.

Tabla 4-12 Medidas características para la evaluación No. 1 de la prueba No. 4.

Medida	English Query	Propuesta
Mejor calificación promedio	79.17 (Considera que la configuración le da una idea de cómo funciona la interfaz)	75.00 (Considera que la configuración de la interfaz es flexible)
Peor calificación promedio	37.50 Considera que el entrenamiento para hacer la tarea de configuración fue el	41.67 (La sesión de entrenamiento le permitió entender la metodología de configuración de la

	adecuado)	interfaz)
Diferencia más alta favorable	-37.50 (Considera que la configuración de la interfaz es inteligible)	16.67 (Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo)
Diferencia más baja favorable	-4.17 (La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz)	8.33 (Considera que la configuración de la interfaz es flexible)

4.5.4.1.2 Análisis de la evaluación No. 1 de la prueba No. 4

- Las calificaciones dadas son bajas, sobre todo para la propuesta, ya que apenas rebasan la mitad de la escala (0 a 100), debido a la poca capacitación que se les dio a los sujetos de prueba.
- La desviación estándar de la pregunta que evalúa los problemas relacionados con el idioma de las ILNBDs es tan alta (35.36 para EQ y 23.57 para la propuesta), que no se contabilizó, debido a que sólo se tomaron en cuenta aquellos promedios cuya desviación para EQ y la propuesta no fueran mayores a 25.
- Las diferencias de los promedios de calificaciones para EQ y la propuesta de aspectos relacionados como flexibilidad (8.33), lo entendible de la interfaz (-37.50) y lo entendible del funcionamiento de la misma (-25.00) son muy extremas.
- La mayoría de las diferencias llega a ser por arriba de 10.

4.5.4.1.3 Conclusiones de la evaluación No. 1 de la prueba No. 4

Conclusiones positivas:

- Los aspectos en que la propuesta sale mejor evaluada que EQ es en el aspecto de flexibilidad, un elemento indispensable para portar de contexto una ILNBD.

Conclusiones negativas:

- El concepto básico de ontologías no fue bien aceptado, debido a que a los evaluadores se les dificulta en general la asimilación de nuevos conceptos,

una habilidad que aún no habían desarrollado por el poco tiempo que llevaban en la maestría.

- En la mayoría de los aspectos evaluados EQ supera a la propuesta, debido a la experiencia previa que tenían los evaluadores con SQL Server, lo que les permitió una mayor familiaridad con EQ que con la propuesta.
- Las diferencias no uniformes en aspectos relacionados indican que no se entendió lo que se estaba evaluando.
- Los comentarios verbales y escritos indican que parte de las malas calificaciones dadas a la propuesta se deben a que los evaluadores hicieron las pruebas contra su voluntad, y bajo una alta carga de estrés, por ser alumnos que iniciaban su maestría.

4.5.4.2 Evaluación No. 2 de la prueba No. 4

Objetivo particular.- Evaluar la asimilación de los conceptos para configurar la interfaz en lenguaje natural hacia bases de datos English Query y otra propuesta de configurar el mismo tipo de interfaces usando ontologías, con consultas de mayor complejidad que las de la evaluación No. 1, los cuales requieren un mayor análisis para diseñar las relaciones que permitan a ambas interfaces responder dichas consultas.

4.5.4.2.1 Resumen de la evaluación No. 2 de la prueba No. 4

En la tabla 4-13 se presentan las diferencias entre cada una de las preguntas que evalúan el proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-13 Diferencias entre aspectos intrínsecos que evalúan el proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la evaluación No. 2 de la prueba No. 4.

Pregunta	Propuesta-EQ
1. Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).	12.50
2. Considera que la configuración de la interfaz es inteligible (que puede ser entendida).	-16.67
3. Considera que la configuración le da una idea de cómo funciona la interfaz (cómo ejecuta las funciones que le son propias).	-25.00

4. Se sintió cómodo al analizar y completar los conceptos de la ontología.	-12.50
5. Se sintió cómodo al analizar y completar las relaciones de la ontología.	-12.50

En la tabla 4-14 se muestran las diferencias entre cada una de las preguntas que describen la evaluación del proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-14 Diferencias entre aspectos extrínsecos que describen la evalución del proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la evaluación No. 2 de la prueba No. 4.

Pregunta	Propuesta-EQ
1. Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado.	8.33
2. La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz.	8.33
3. Me sentí cómodo con el ambiente de trabajo después de la sesión de entrenamiento.	-8.33
4. Fue sencillo aprender a usar el editor Protégé o el ambiente de configuración de English Query (según sea el caso).	-4.17
5. Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo.	25.00
6. Considera que la documentación de la configuración de la interfaz es fácil de entender (tener idea clara de las cosas).	-12.50
7. Considera que la terminología usada para configurar la interfaz es extraña o confusa.	-25.00

Nota: un valor positivo en la columna Propuesta-EQ, significa que se consideró mejor la propuesta del uso de ontologías y un valor negativo significa que se consideró mejor a English Query.

Tabla 4-15 Medidas características para la evaluación No. 2 de la prueba No. 4.

Medida	English Query	Propuesta
Mejor calificación promedio	79.17 (Considera que la configuración le da una idea de cómo funciona la	79.17 (Afecta en la configuración el que en la interfaz haya elementos que no

	interfaz)	son en su idioma nativo)
Peor calificación promedio	50.00 (Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado)	54.17 (varias)
Diferencia más alta favorable	-25.00 (Considera que la configuración le da una idea de cómo funciona la interfaz)	25.00 (Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo)
Diferencia más baja favorable	-4.17 (Fue sencillo aprender a usar el ambiente de configuración de English Query)	8.33 (Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado)

4.5.4.2.2 Análisis de la evaluación No. 2 de la prueba No. 4

- La mayor parte de los resultados de los promedios normalizados mejoran notablemente tanto para EQ como para la propuesta.
- En las preguntas de la evaluación de la comodidad con el ambiente de trabajo y el entendimiento de la metodología, la propuesta sale mejor evaluada que EQ.
- Los aspectos peor evaluados tienen que ver con la capacitación recibida.
- Aunque la propuesta sale mejor evaluada con respecto a los problemas del lenguaje usado (inglés), no es posible compararla porque la desviación estándar del promedio de la evaluación para EQ es muy alta (37.96).

4.5.4.2.3 Conclusiones de la evaluación No. 2 de la prueba No. 4

Conclusiones positivas:

- La propuesta vuelve a salir mejor evaluada en flexibilidad.
- El concepto básico de ontología fue mejor aceptado que en la evaluación anterior.

- Se nota en general un aumento en las calificaciones sobre todo para la propuesta. A pesar de la poca capacitación que se les dio a los sujetos de prueba, la experiencia de la primera evaluación les ayudó a comprender mejor el proceso de configuración, en especial el proceso propuesto que involucra ontologías.
- El aumento en la complejidad de las consultas no parece afectar a este grupo de prueba.

Conclusiones negativas:

- De doce preguntas que se presentan, la propuesta sólo es mejor en cuatro y EQ en ocho.
- El hecho de que se considere más entendible el funcionamiento de EQ, es debido a que en las evaluaciones de la prueba No. 4 se les permitió a los sujetos de prueba probar las respuestas de EQ, algo que no es factible con la propuesta.

4.5.4.3 Evaluación No. 3 de la prueba No. 4

Objetivo particular.- Evaluar la asimilación de los conceptos para configurar interfaces de lenguaje natural hacia bases de datos, en la interfaz comercial English Query y en una propuesta usando ontologías, con consultas de igual complejidad que las de la evaluación No 1 de la prueba No. 4, pero con la base de datos *Pubs*, en lugar de la base de datos *Northwind*, la cual fue usada en las evaluaciones No. 1 y No. 2.

4.5.4.3.1 Resumen de la evaluación No. 3 de la prueba No. 4

En la tabla 4-16 se presentan las diferencias entre cada una de las preguntas que evalúan el proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-16 Diferencias entre aspectos intrínsecos que evalúan el proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la evaluación No. 3 de la prueba No. 4.

Pregunta	Propuesta-EQ
1. Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).	16.67
2. Considera que la configuración de la interfaz es inteligible (que puede ser entendida).	-16.67
3. Considera que la configuración le da una idea de cómo	-16.67

funciona la interfaz (cómo ejecuta las funciones que le son propias).	
4. Se sintió cómodo al analizar y completar los conceptos de la ontología.	0.00
5. Se sintió cómodo al analizar y completar las relaciones de la ontología.	12.50

En la tabla 4-17 se muestran las diferencias entre cada una de las preguntas que describen la evaluación del proceso de configuración propuesto basado en una ontología contra el proceso que implica configurar English Query.

Tabla 4-17 Diferencias entre aspectos extrínsecos que describen la evalución del proceso de configuración basado en una ontología contra el proceso de configuración de English Query para la evaluación No. 3 de la prueba No. 4.

Pregunta	Propuesta-EQ
1. Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado.	8.33
2. La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz.	12.50
3. Me sentí cómodo con el ambiente de trabajo después de la sesión de entrenamiento.	-4.17
4. Fue sencillo aprender a usar el editor Protégé o el ambiente de configuración de English Query (según sea el caso).	-4.17
5. Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo.	25.00
6. Considera que la documentación de la configuración de la interfaz es fácil de entender (tener idea clara de las cosas).	-12.50
7. Considera que la terminología usada para configurar la interfaz es extraña o confusa.	4.17

Nota: un valor positivo en la columna Propuesta-EQ, significa que se consideró mejor la propuesta del uso de ontologías y un valor negativo significa que se consideró mejor a English Query.

Tabla 4-18 Medidas características para la evaluación No. 3 de la prueba No. 4.

Medida	English Query	Propuesta
Mejor calificación promedio	79.17 (Fue sencillo aprender a usar el editor Protégé o el ambiente de English Query)	91.67 (Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo)
Peor calificación promedio	58.33 (La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz)	54.17 (Considera que la configuración le da una idea de cómo funciona la interfaz)
Diferencia más alta favorable	-16.67 (Considera que la configuración le da una idea de cómo funciona la interfaz)	25.00 (Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo)
Diferencia más baja favorable	0 (Se sintió cómodo al analizar y completar los conceptos de la ontología)	0 (Se sintió cómodo al analizar y completar los conceptos de la ontología)

4.5.4.3.2 Análisis de la evaluación No. 3 de la prueba No. 4

- La mayor parte de los resultados de los promedios normalizados mejoran en la evaluación de la interfaz, con un incremento promedio de 3.33 para EQ y de 10.83 para la propuesta.
- Se mantiene mejor evaluada la propuesta que EQ en las preguntas de la evaluación del entrenamiento relacionadas con la comodidad con el ambiente de trabajo y el entendimiento de la metodología.
- En las preguntas de la evaluación de la metodología relacionadas con claridad y comodidad del proceso de configuración, se consideró mejor la propuesta; mientras en el análisis de los conceptos, EQ y la propuesta salen igualmente evaluadas.
- Los aspectos peor evaluados para la propuesta tienen que ver con el entendimiento del funcionamiento de la interfaz.

- Aunque la propuesta sale mejor evaluada con respecto a los problemas del lenguaje, no es posible compararla por la desviación estándar del promedio de la evaluación para EQ (26.35).
- Otro aspecto en el que vuelve a salir mejor evaluada la propuesta es flexibilidad.
- De las doce preguntas presentadas, la propuesta es mejor en seis, EQ mejor en cinco, y en una sus resultados son iguales.

4.5.4.3.3 Conclusiones de la evaluación No. 3 de la prueba No. 4

Conclusiones positivas:

- La propuesta vuelve a salir mejor evaluada en flexibilidad, igual que en la evaluación No. 1 y No. 2, siendo un aspecto fundamental en el propósito de hacer portable una ILNBD.
- Se nota en general un aumento en las calificaciones, sobre todo para la propuesta. A pesar de la poca capacitación que se les dio a los sujetos de prueba, la experiencia de dos evaluaciones anteriores les ayudó a comprender mejor el proceso de configuración, en especial el proceso propuesto que involucra ontologías. Este mejor entendimiento del concepto de ontologías se refleja en que la propuesta sale mejor evaluada en general en los dos grupos de aspectos evaluados.
- La disminución en la complejidad de las consultas hace que mejoren las calificaciones.
- En general la propuesta sale un poco mejor evaluada que EQ.

Conclusiones negativas:

- Aunque en general la propuesta sale mejor evaluada, su ventaja no supera en promedio el 2%, siendo las diferencias más bien compensatorias, es decir, si sale mal evaluada la propuesta en un aspecto, sale bien evaluada en otro.
- El hecho de que aspectos relacionados tengan calificaciones contrarias en cada evaluación, indican que varios aspectos de las pruebas no fueron completamente comprendidos por los sujetos de prueba.

4.5.4.4 Conclusiones generales de la prueba No. 4

En esta sección se compararán los resultados de las tres evaluaciones de la prueba No. 4, tanto para EQ como para la propuesta de utilizar ontologías para configurar una ILNBD.

4.5.4.4.1 Evolución de los promedios de las calificaciones que evalúan los procesos de configuración en las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4

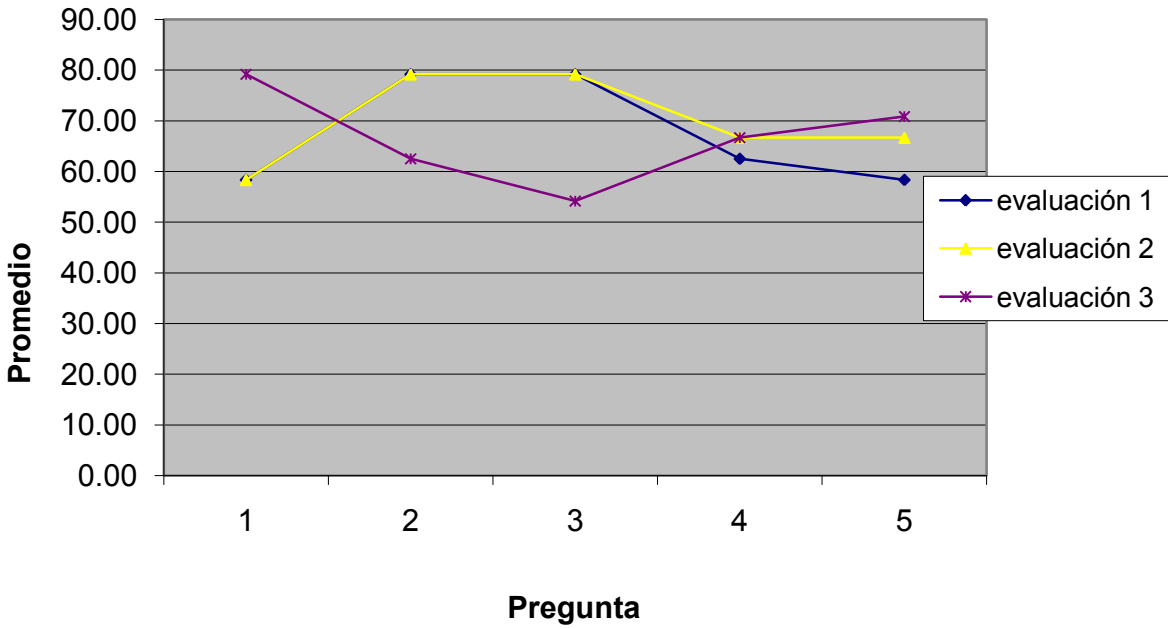

Figura 4-1 Gráfica de la evolución de los promedios de las calificaciones de aspectos intrínsecos que evalúan el proceso de configuración de EQ para las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4.

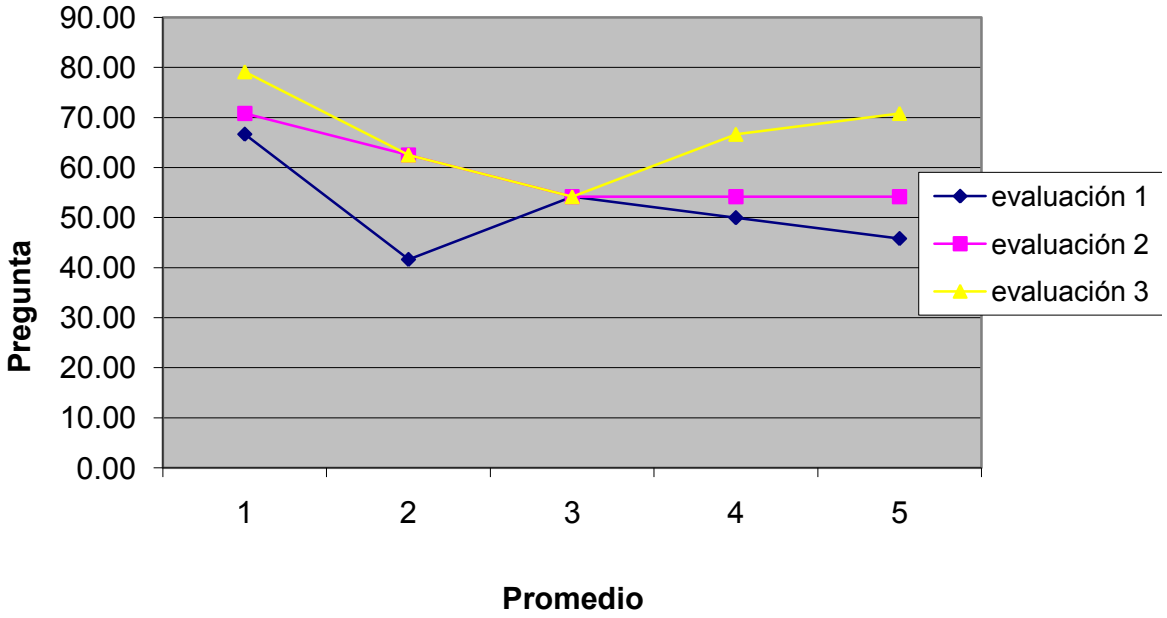

Figura 4-2 Gráfica de la evolución de los promedios de las calificaciones de aspectos intrínsecos que evalúan el proceso de configuración basado en ontologías para las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4.

Conclusiones positivas:

- El comportamiento de los promedios para la propuesta fue casi uniforme, a diferencia de los promedios de EQ que varían negativamente a través de las pruebas (figuras 4-1 y 4-2).
- El factor del aumento en la complejidad de las consultas a configurar es lo que hace la diferencia entre la segunda evaluación y las otras dos, por lo que se aprecia que la propuesta es menos susceptible a la complejidad que EQ.
- La experiencia adquirida en cada evaluación se manifiesta mucho más en la propuesta que en EQ. Un aspecto que afectó en las calificaciones relacionadas con la interfaz fue que la ontología predefinida para configurar estaba en español, aunque las herramientas usadas, Protégé y el ambiente de configuración de English Query, están en inglés.

Conclusiones negativas:

- A pesar de que el comportamiento de los promedios dados a la propuesta mejora en cada evaluación, no se tiene una mejoría en todos los aspectos, lo cual es coherente con la percepción de que EQ es una ILNBD totalmente funcional y la propuesta no.

4.5.4.4.2 Factores extrínsecos de las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4

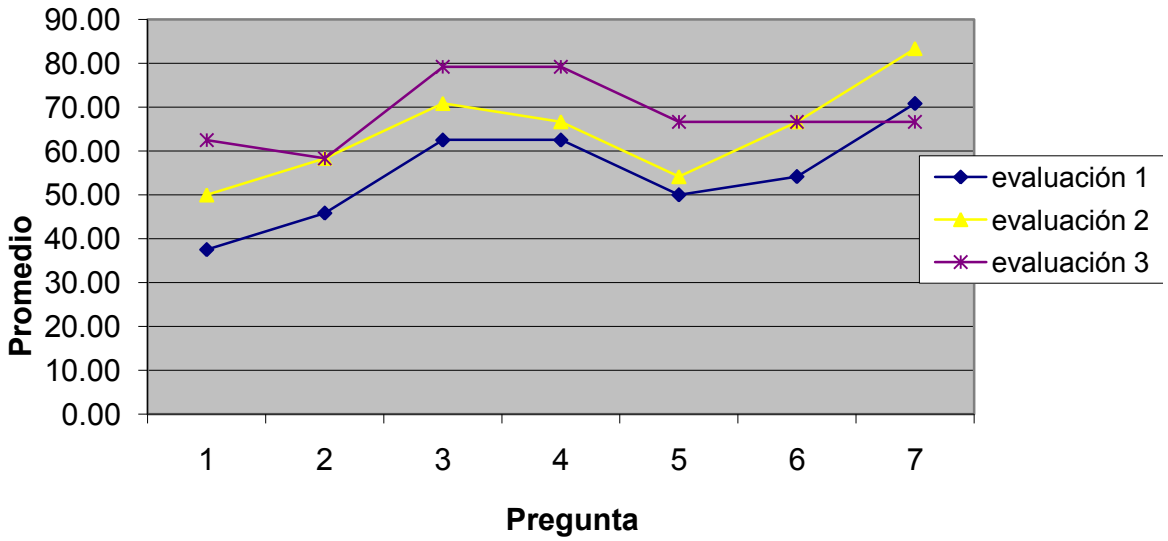

Figura 4-3 Gráfica de la evolución de los promedios de las calificaciones para los factores extrínsecos de las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4 para EQ.

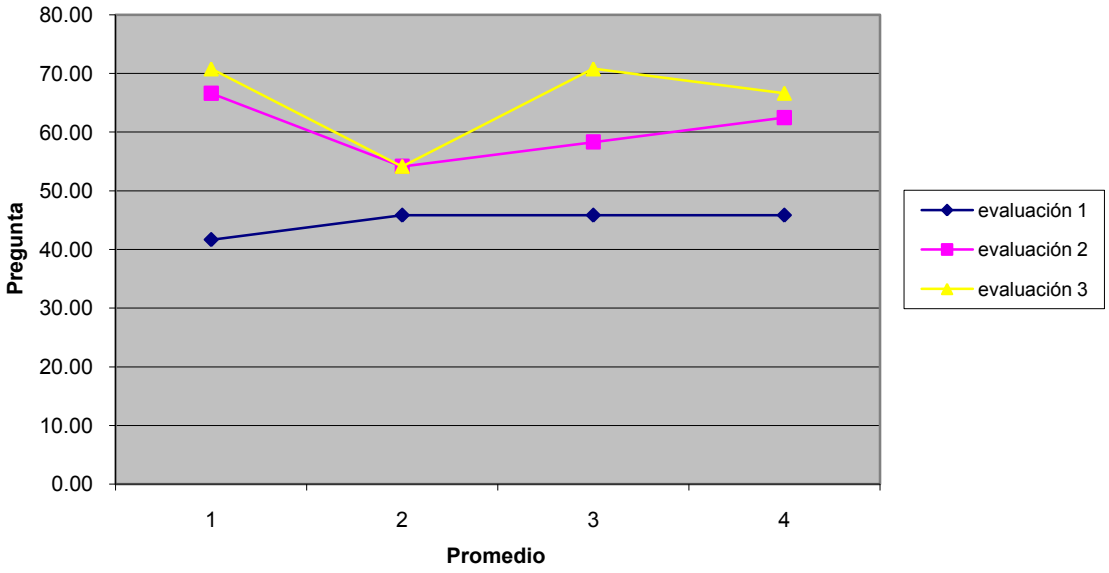

Figura 4-4 Gráfica de la evolución de los promedios de las calificaciones para los factores extrínsecos de las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4 para la propuesta.

Conclusiones positivas:

- La propuesta presenta una metodología más completa que la de EQ, lo cual se refleja en el hecho de que los promedios dados a la propuesta no presentan tantos altibajos (figura 4-4), como los que presenta EQ (figura 4-3).
- La propuesta tiene mejoras significativas entre la primera y la segunda evaluación, y mejora menor entre la segunda y tercera. La razón principal es que la propuesta es mejor para configurar las respuestas de consultas complejas, pero no lo es para consultas sencillas, sobre todo porque las consultas sencillas pueden ser contestables por una configuración generada de manera automática.

Conclusiones negativas:

- Los promedios de calificaciones, en especial los de la propuesta, son muy bajos en la evaluación No. 1 (figura 4-4).
- Los promedios relacionados con calificaciones en EQ tienen un altibajo (pregunta 7) y estancamientos (preguntas 2 y 6), posiblemente asociados a que la complejidad de las evaluaciones No. 1 y No. 3 son la misma y la complejidad de la evaluación No. 2 es mayor, aunque no se encontró un patrón claro para estos altibajos (figura 4-3).

4.5.4.4.2 Evaluación de las diferencias de las calificaciones en las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4

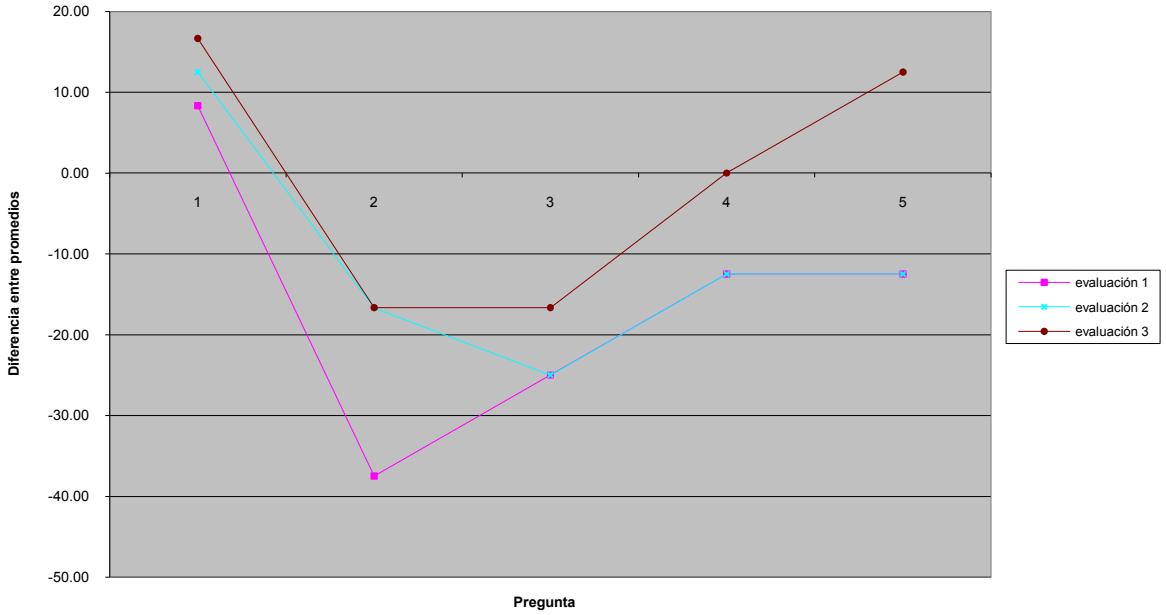

Figura 4-5 Gráfica de la evolución de las diferencias de los promedios de las calificaciones de aspectos intrínsecos que evalúan la propuesta de configuración basada en ontologías contra el proceso de configuración de EQ en las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4 (una diferencia positiva indica que la propuesta se consideró mejor y una diferencia negativa lo contrario).

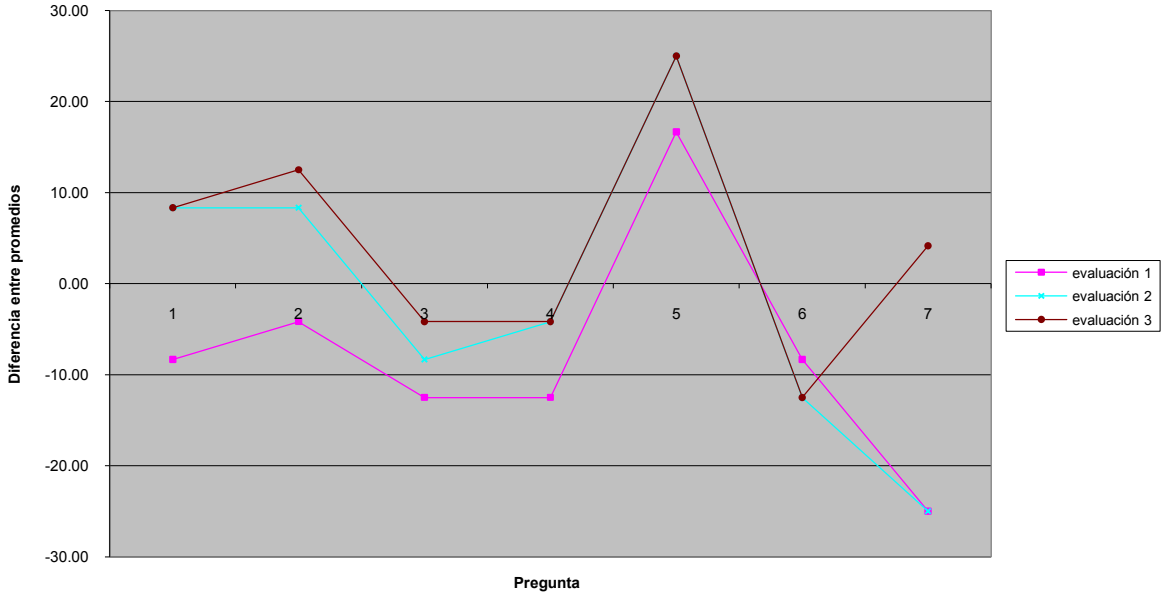

Figura 4-6 Gráfica de la evolución de las diferencias de los promedios de las calificaciones de aspectos extrínsecos de los factores que describen la evaluación de la propuesta de configuración con ontologías contra el proceso de configuración de EQ en las evaluaciones No. 1, No. 2 y No. 3 de la prueba No. 4 (una diferencia positiva indica que la propuesta se consideró mejor y una diferencia negativa lo contrario).

Conclusiones positivas:

- A pesar de que en los aspectos relacionados con la interfaz, la propuesta tenía casi todas sus evaluaciones en contra (valores negativos), se percibe en las siguientes dos evaluaciones una tendencia a la alza (figura 4-5).
- Se tienen sólo dos altibajos en la propuesta y están asociados a lo entendible de la documentación y a la terminología usada. Estos altibajos están asociados a los problemas que presentan los sujetos de prueba para asimilar los conceptos de ontologías de una manera autodidacta (figura 4-6).

Conclusiones negativas:

- Se notan pocos cambios en las diferencias de las preguntas que evalúan los mecanismos de configuración para las evaluaciones No. 2 y No. 3 (figura 4-5).
- A pesar de las innegables mejoras en cuanto a los promedios de calificaciones de la propuesta que se tienen en cada evaluación, cinco de los doce aspectos presentados aún terminan a favor de EQ y uno empatado en la tercera evaluación.

4.5.5 Conclusiones generales del plan de pruebas

El objetivo de las pruebas es comprobar la validez de la hipótesis planteada, a través de la comparación de las calificaciones dadas a EQ y las dadas a la propuesta de ontologías para configurar una ILNBD. Se muestra primero una comparación entre las pruebas No. 2 y No 3, debido a que fueron iguales en cuanto a los aspectos a evaluar, y al final de esta sección se muestra una comparación entre los aspectos comunes a las pruebas No. 1, No. 2 y No. 3.

4.5.5.1 Evaluación de las diferencias entre la propuesta y EQ en las pruebas No. 2 y No. 3

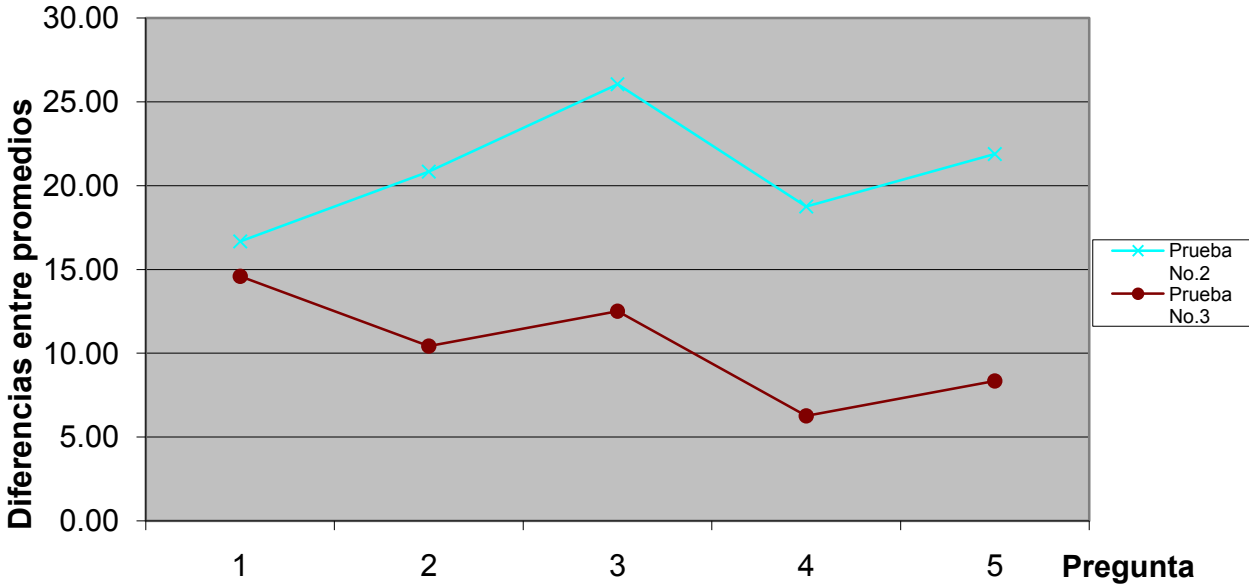

Figura 4-7 Gráfica de las diferencias de los promedios de las calificaciones de aspectos intrínsecos que evalúan los procesos de configuración de EQ contra el proceso basado en una ontología en las pruebas No. 2 y prueba No. 3 (una diferencia positiva indica que la propuesta se consideró mejor, y una diferencia negativa, lo contrario).

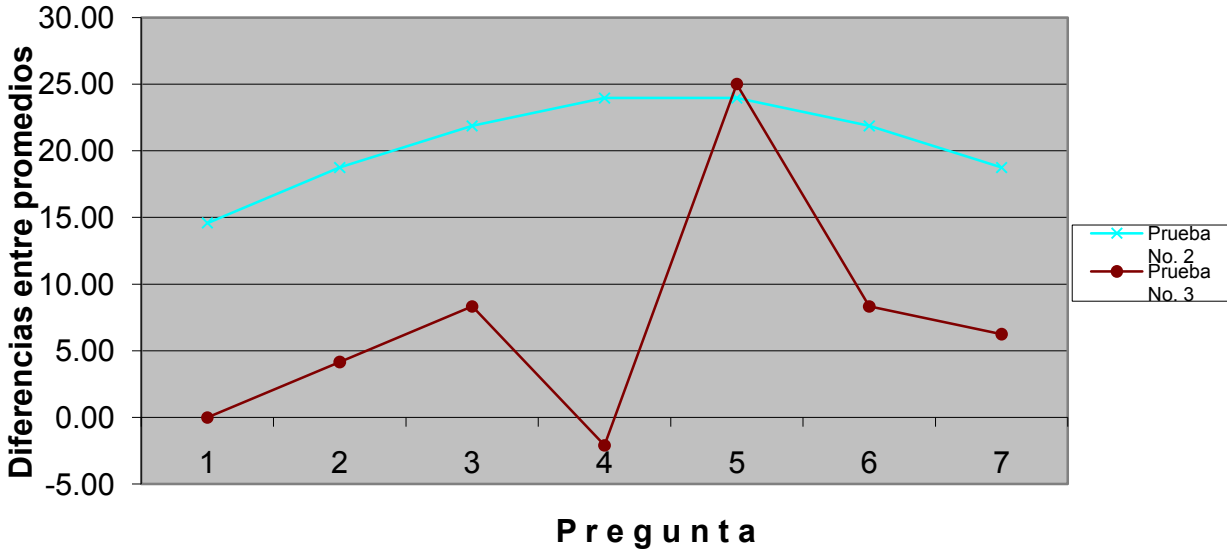

Figura 4-8 Gráfica de las diferencias de los promedios de las calificaciones de aspectos intrínsecos que evalúan los procesos de configuración de EQ contra el proceso basado en una ontología en las pruebas No. 2 y prueba No. 3 (una diferencia positiva indica que la propuesta fue mejor, y una diferencia negativa, lo contrario).

Conclusiones positivas:

- La mayoría de las diferencias están a favor de la propuesta: la propuesta sale mejor evaluada en todos los aspectos que evalúan los procesos de configuración; mientras que en los factores extrínsecos, únicamente en la prueba No. 3 EQ sale mejor evaluada en la pregunta No. 4 “Fue sencillo aprender a usar el editor Protégé o el ambiente de configuración de English Query” debido a que los participantes de esta prueba están más familiarizados con las interfaces comerciales.
- Las diferencias más contrastantes se dan en los siguientes puntos: pregunta No. 4 de los factores extrínsecos “Fue sencillo aprender a usar el editor Protégé o el ambiente de English Query” (de 23.96 en la prueba No. 2 a -2.08 en la prueba No. 3, figura 4-8), y “Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado“(de 15 en la prueba No. 2 a 0 en la prueba No. 3, figura 4-8).
- Las razones por las que se dan las diferencias anteriores podrían ser debido a que los evaluadores de la prueba No. 2 están más acostumbrados a ser autodidactas.
- Las diferencias a favor de la flexibilidad (pregunta No. 1) y de la claridad con que se entiende el funcionamiento de la ILNBD (pregunta No. 3) continúan siendo las más favorables a la propuesta para los aspectos intrínsecos (figura 4-7).

Conclusiones negativas:

- En la prueba No. 3 se presentan más diferencias a favor de EQ, siendo las posibles razones que los participantes de la prueba No. 3 tienen más experiencia laboral, pero menos capacidades para analizar.

4.5.5.2 Evaluación de las preguntas comunes de la propuesta y EQ en las pruebas No. 1, No. 2 y No. 3

Las preguntas comunes a las pruebas No. 1, No. 2 y No. 3 son las siguientes:

1. Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).
2. Considera que la configuración de la interfaz es manejable (moverse con cierta soltura después de haber tenido algún impedimento).
3. Considera que la configuración de la interfaz es inteligible (que puede ser entendida).

4. Considera que la configuración le da una idea de cómo funciona la interfaz (cómo ejecuta las funciones que le son propias).
5. Considera que la terminología usada para configurar la interfaz es extraña o confusa.
6. Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo

Aclaración.- El orden original en que vienen estas preguntas para las pruebas No. 1, No. 2 y No. 3 varía, por lo cual, se reordenaron para su comparación.

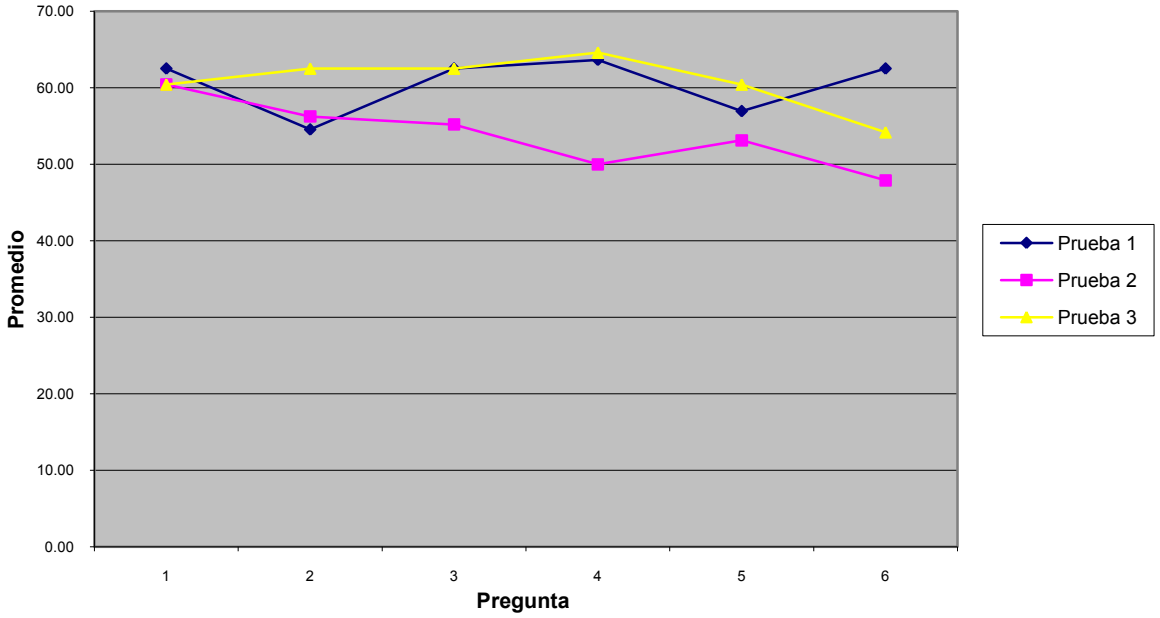

Figura 4-9 Gráfica de la evolución de las preguntas comunes en las pruebas No. 1, No. 2 y No. 3 para EQ.

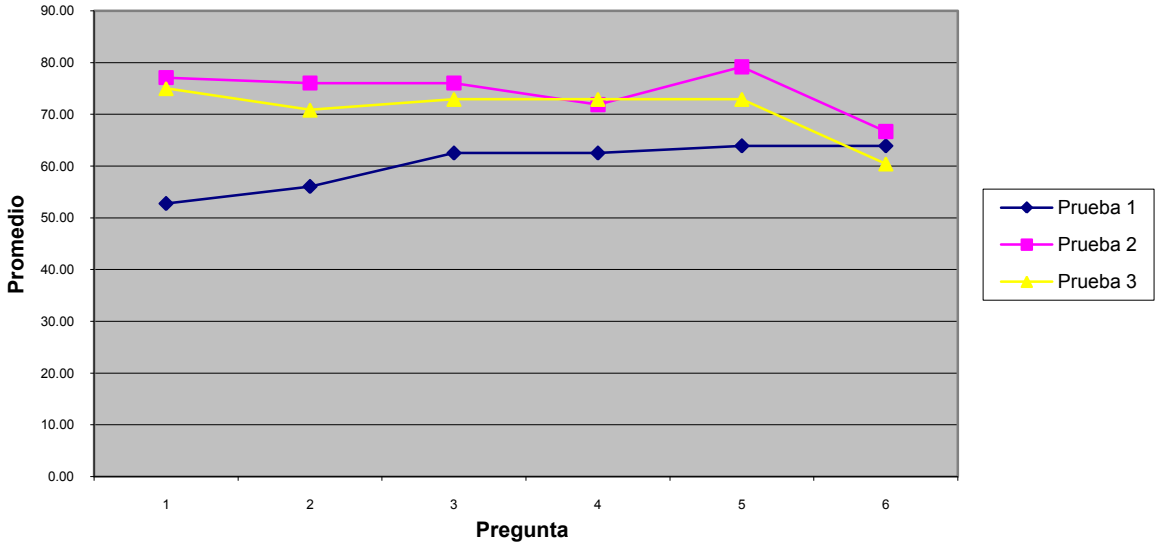

Figura 4-10 Gráfica de la evolución de las preguntas comunes en las pruebas No. 1, No. 2 y No. 3 para la propuesta.

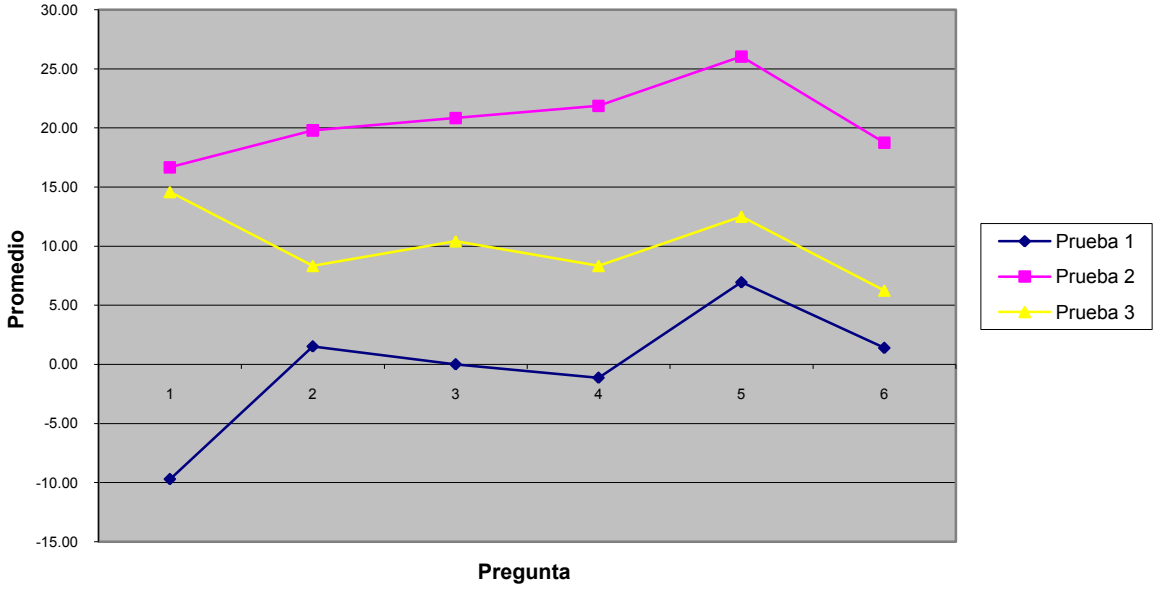

Figura 4-11 Gráfica de la evolución de las diferencias entre las preguntas comunes para la propuesta y EQ en las pruebas No. 1, No. 2 y No. 3 (una diferencia positiva indica que la propuesta se consideró mejor, y una diferencia negativa, lo contrario).

Conclusiones positivas:

- Las diferencias que se dieron entre las pruebas No. 1, No. 2 y No. 3 (mejoras en la documentación, simplificación de la ontología en la prueba de afinación, diferente tipo de sujetos de prueba, etcétera) harían suponer que la comparación de los aspectos comunes a las tres pruebas

(flexibilidad, facilidad de uso, manejabilidad, etcétera) nos daría una gráfica en donde las calificaciones de la prueba No. 2, fueran superiores a las de las prueba No. 3 (porque los sujetos de la prueba No. 2 son alumnos de Cenidet; y por lo tanto, se presume que deberían tener una mayor capacidad de análisis) y a su vez las calificaciones de la prueba No. 3 deberían ser mejores que las de la prueba No. 1; sin embargo, para la propuesta hay un cruce en la pregunta 6 (figura 4-10), mientras que para EQ hay cruce en casi todas las preguntas (figura 4-9).

- El único aspecto que cambió significativamente entre las pruebas No. 1, No. 2 y No. 3 fue el incremento y mejoramiento de la documentación, lo cual se refleja en las diferencias entre las pruebas (figura 4-11).
- En la prueba No. 1 se tienen diferencias negativas, mientras que a partir de la prueba No. 2 todas las diferencias son positivas, es decir, a favor de la propuesta (figura 4-11).
- La pregunta 1, relacionada con la flexibilidad, muestra el mayor cambio entre la prueba No. 1 y las pruebas subsecuentes (figura 4-11). En general todas las preguntas comunes tienen mejoras significativas entre prueba y prueba.

Conclusiones negativas:

- Los valores más bajos se dan para la prueba No. 1, debido a que no todos los participantes la terminaron; por lo tanto, los resultados de la misma no son muy confiables.

CAPITULO 5
Conclusiones

En este capítulo se presentan las conclusiones obtenidas a través de todo el proceso de investigación desarrollado para llevar a cabo esta tesis, las

aportaciones más relevantes que surgieron a través de mismo, y algunos posibles trabajos futuros que la pueden complementar.

5.1 Validación de las hipótesis propuestas

Se muestran a continuación las hipótesis planteadas en este trabajo:

H1. *El uso de una ontología para representar el conocimiento lingüístico, reduce el esfuerzo de configurar una ILNBD para diferentes tipos de consultas.*

H2. *El uso de una ontología para representar el conocimiento lingüístico, reduce el esfuerzo de portar una ILNBD para diferentes dominios.*

Los aspectos comunes de las pruebas No. 1, No. 2 y No. 3 están relacionados con el concepto de facilidad en general, así que el comportamiento de mejoría que presentan permite afirmar que la propuesta del uso de ontologías permite configurar una ILNBD de una mejor manera que EQ y comprobar la hipótesis H1.

La flexibilidad fue un aspecto relevante en la propuesta desde la prueba No. 1 hasta la prueba No. 4. Considerando que este aspecto es primordial para lograr la portabilidad de dominio de una ILNBD, podría afirmarse que se comprueba la hipótesis H2.

En las pruebas No. 4 y de afinación, algo destacable se descubrió al comprobar que los sujetos de prueba resienten mucho más el cambio en la complejidad de las consultas que el cambio de base de datos.

5.2 Aportaciones

- Una metodología de configuración más flexible y más expresiva que las encontradas en otras ILNBDs, ya que permite definir la información necesaria para contestar una mayor diversidad de consultas y de una manera más clara para el personalizador de la interfaz.
- Una ontología de propósito general que incorpora elementos que conforman una base de datos relacional, y que da soporte a la metodología como enlace entre las palabras que constituyen una consulta con los elementos de la base de datos, que le serán útiles al analizador semántico para construir la consulta en SQL.
- Una evaluación cuantitativa del proceso de configuración de una ILNBD desde el punto de vista del configurador, algo no encontrado en la revisión de la literatura especializada.

5.3 Conclusiones

No se encontraron en la literatura especializada evaluaciones del proceso de configuración de las ILNBDs, por lo que este trabajo es pionero en su campo. Aunque existe mucho trabajo e interés en aspectos de usabilidad para el diseño de interfaces de usuario, el proceso de configuración de una base de conocimiento es diferente, ya que involucra, además de ciertas tareas repetitivas, actividades que implican cierto conocimiento del funcionamiento de la aplicación y, para el caso particular de las ILNBDs, ciertos conocimientos de lingüística.

A pesar de que EQ es una ILNBD completa y la propuesta no, fue más deseable para los sujetos de prueba conocer todos los términos y sus relaciones, es decir, una base de conocimiento explícita (ontología), que los elementos de soporte (ayuda en línea, editor gráfico de relaciones, transparencia en el proceso de traducción, etc.).

Los resultados de las pruebas No. 4 y de evaluación llevan a concluir que más que comprobar la hipótesis H2, la barrera a superar no es el que la ILNBD trabaje con diferentes bases de datos, sino que trabaje con diferentes tipos de consultas. Lo anterior puede ayudar a definir una medida del trabajo que implica traducir la consulta en lenguaje natural al lenguaje SQL.

5.4 Trabajos futuros

Este trabajo de tesis puede ser continuado en otros proyectos a realizar:

- *Traductor de lenguaje natural a SQL basado en ontologías.* Un complemento muy importante de esta propuesta sería el desarrollo de un traductor semejante al de [12], pero que aprovechara totalmente la configuración definida en la ontología propuesta, para la traducción de la consulta en lenguaje natural hecha por el usuario final, al lenguaje SQL.
- *Mecanismo de mezcla y optimización de ontologías.* Un aspecto sin explotar en este trabajo fue la mezcla de ontologías para disminuir el trabajo del personalizador, así como la optimización de dicha mezcla, con el fin de ajustar el contenido de la ontología a las necesidades de la ILNBD para el contexto de la base de datos en la que se esté trabajando.
- *Mecanismo de aprendizaje de ontologías extendido.* El mecanismo de adquisición de ejemplares de la ontología predefinida sólo abarcó las relaciones *es un* y *parte de*, por lo que un trabajo complementario sería el

estudio de otro tipo de relaciones (actor, instrumento, pertenencia, espacial, etc.).

- *Asistente para la configuración de una interfaz en lenguaje natural.* Aunque se desarrolló un editor de ontologías [8] para asistir al configurador a personalizar la ontología de la ILNBD para una determinada base de datos, no se terminó, y sería conveniente terminar el editor y complementarlo con algunas características de herramientas semejantes de English Query y ELF.

5.5 Publicaciones y proyectos

Proyectos apoyados

- *Ontología del Español para Aplicaciones del Lenguaje*, proyecto COSNET, clave 877.03-P.
- *Interfaz en Lenguaje Natural hacia Bases de Datos para Usuarios de Internet*, proyecto COSNET, clave 571.01-P.
- *Interfaz en Lenguaje Natural hacia Bases de Datos para Usuarios de Internet*, apoyo CYTED a la Red Iberoamericana de Tecnologías del Software RITOS-II.

Artículos en Lecture Notes in Computer Science:

- Jose A. Zárate M., Rodolfo A. Pazos R., Alexander Gelbukh, Joaquin Pérez O., "Improving the Customization of Natural Language Interface to Databases Using an Ontology", *Proc. of International Conference on Computational Science and Its Applications (ICCSA 2007)*, LNCS 4705, Kuala Lumpur, Ago. 2007, ISSN 0302-9743, pp. 424-435.
- J. Antonio Zárate M., Rodolfo A. Pazos R., Alexander Gelbukh e J. Isabel Padrón C., "A Portable Natural Language Interface for Diverse Databases Using Ontologies", *Proc. of Computational Linguistics and Intelligent Text Processing (CiCLing 2003),* LNCS 2588, Ciudad de México, Feb. 2003, ISSN 0302-9743, pp. 494-505.

Artículos en congresos internacionales:

- Jose A. Zárate, Rodolfo A. Pazos R. y Alexander Gelbukh, "Customization of Natural Language Interfaces to Databases: Beyond Domain Portability", *Memoria del Magno Congreso del CIC* a publicarse por la IEEE Computer Society, Nov. 2007.

- Jose A. Zárate, Rodolfo A. Pazos, Roberto Toledo, "Acquisition of Lexical-Syntactic Relationships from a Dictionary", *Memoria del 11vo. Congreso Internacional de Investigación en Ciencias Computacionales*, Tlalnepantla, México, Sept. 2004, ISBN 968-5823-10-3, pp. 45-52.
- Roberto Toledo, José A. Zárate, Rodolfo A. Pazos, "Ontología del Español para una Interfaz en Lenguaje Natural hacia Bases de Datos"*, Memoria del 10mo. Congreso Internacional de Investigación en Ciencias Computacionales*, Oaxtepec, México, Oct. 2003, ISBN 968-5823-02-2, pp. 56-61.
- José A. Zárate, Rodolfo A. Pazos, Alexander Gelbukh, "Natural Language Interface for Web-based Databases", *Proc. of 2nd WSEAS int. conf. on Robotics, Distance Learning and Intelligent Communication Systems (ICRODIC 2002)*, Skiathos, Grecia, Oct. 2002, ISBN 906-8052-68-8.
- José A. Zárate, Rodolfo A. Pazos., Alexander Gelboukh K.*, "*Interfaz en Lenguaje Natural para Usuarios de Bases de Datos en Internet"*, Memoria del V Worshop Latinoamericano de Tecnologías del Software, IDEAS`02,* La Habana*,* Cuba, Abr. 2002, pp. 401-404.

Tesis de maestría terminadas derivadas de este trabajo:

- Roberto Carlos Toledo Flandes*, Generación de una Ontología de Dominio Lingüístico para el Español,* tesis de maestría, CENIDET, Feb. 2005.

ANEXO A: **Diseño de la ontología predefinida**

Se muestra el modelo conceptual con el que se implementó una ontología genérica a utilizar como base de conocimiento para una interfaz en lenguaje natural hacia bases de datos (ILNBD). Para tal efecto, se describen los conceptos utilizados y las relaciones en las que participan. Se presentan en general sólo los niveles superiores de la ontología, por cuestiones de espacio.

Clases o categorías principales definidas en la ontología.

Concepto:	**Palabra.**
Definición:	(Del lat. *parabŏla*). **1.** f. Segmento del discurso unificado habitualmente por el acento, el significado y pausas potenciales inicial y final. **2.** f. Representación gráfica de la **palabra** hablada. Dado que "palabra" normalmente se refiere a la pronunciación y a su concepto asociado, las discusiones de esta asociación léxica son vulnerables a una confusión terminológica. Con el ánimo de reducir ambigüedad, aquí se usará *significante* (word form) para referirse a la pronunciación física o la escritura, y *significado* (word meaning) se usará para referirse al concepto lexicalizado que un significante puede usar para expresar algo.
Superclase:	Raíz (Cosa).
Ejemplo(s):	*Escuela* (significante), *institución dedicada a la enseñanza* (significado).

Concepto:	**Synset.**
Definición:	Agrupación de significantes que comparten un significado (es decir, están unidos por la relación de sinonimia) y que se refiere a una clase (que consta de varias entidades semejantes). Este conjunto hereda los

	atributos de los significantes que agrupa y puede ser identificado por uno de sus significantes.
Superclase:	Raíz (Cosa).
Ejemplo(s):	*Tomar1*, *casa1-casa*, *hogar*, *domicilio*, *residencia*. El significante *español* en su acepción de idioma, no es un synset ya que es un ejemplar del significante *idioma*. El significante *azul* tampoco es synset porque es un ejemplar de *color*.

Concepto:	**Sustantivo.**
Definición:	Clase o categoría referente a entidades que tienen existencia real, independiente y/o individual. Clase de palabras que pueden funcionar como sujeto de la oración. También se usa para referirse a un significante que pertenece a esta clase o categoría.
Superclase:	Palabra.
Subclases:	*Concepto*, *ser viviente* y *objeto físico*.

Concepto:	**Verbo.**
Definición:	Clase o categoría de significantes que expresan una idea o acción y que tienen variación de tiempo, modo, persona y número. También se usa para referirse a un significante que pertenece a esta clase o categoría.
Superclase:	Palabra.
Subclases:	*Cambio*, *comunicación*, *consumo*, *competencia*, *relacionados con el clima*, *contacto*, *cognición*, *creación*, *movimiento*, *relativo al cuidado del cuerpo*, *situación*, *posesión*, *interacción social*, *percepción* y *sentimiento*.

Concepto:	**Adjetivo.**
Definición:	Clase o categoría de significantes que califican o determinan al sustantivo. También se usa para referirse a un

	significante que pertenece a esta clase o categoría.
Superclase:	Palabra.
Subclases:	*Demostrativo*, *posesivo*, *indefinido*, *numeral* e *interrogativo*.

Concepto:	**Adverbio.**
Definición:	Clase o categoría de significantes cuya función consiste en complementar el significado del verbo, un adjetivo u otro adverbio, y que tienen la característica de ser invariables. También se usa para referirse a un significante que pertenece a esta clase o categoría.
Superclase:	Palabra.
Subclases:	*Negación*, *duda*, *afirmación*, *modo*, *cantidad*, *tiempo* y *lugar*.

Concepto:	**Otros.**
Definición:	Clase o categoría de significantes que no pertenecen a ninguna de las siguientes clases o categorías: sustantivo, verbo, adjetivo y adverbio.
Superclase:	Palabra.
Subclases:	*Pronombre*, *preposición*, *artículo* y *conjunción*.

Concepto:	**ElementosBD.**
Definición:	Clase o categoría de identificadores de objetos de base de datos definidos en el modelo relacional.
Superclase:	Raíz (Cosa).
Subclases:	*Tabla*, *columna*, *base_datos* y *dominio*.

Concepto:	**Funciones.**
Definición:	Clase o categoría de identificadores de funciones que servirán como enlace

	hacia los programas y funciones que extienden la funcionalidad de la interfaz.
Superclase:	Raíz (cosa).
Subclases:	*Funciones definidas por el usuario, relaciones definidas por el usuario, funciones de agregación.*

Concepto:	**Referencia.**
Definición:	Clase o categoría de identificadores que sirven como parámetros de referencia.
Superclase:	Raíz (cosa).
Subclases:	*Estado_local, ciudad_local* y *dia_hoy.*

Relaciones y propiedades principales definidas en la ontología.

Relación:	**Relación léxica.**
Definición:	Un modelo culturalmente reconocido de asociación que existe entre las unidades léxicas en un idioma.
Cardinalidad:	Definida en sus subrelaciones.
Estructura adyacente:	Definida en sus subrelaciones.
Dominio:	Definido en sus subrelaciones.
Rango:	Definido en sus subrelaciones.
Propiedades:	Definidas en sus subrelaciones.
Superclase:	Relación o propiedad de OWL.
Subclases:	*Relación léxica paradigmática* y *relación léxica sintagmática*.

Relación:	**Relación léxica sintagmática.**
Definición:	Relación léxica entre pares de unidades léxicas (A1-B1, A2-B2, A3-B3,..) donde los miembros de cada par (A1 y B1)

	cumplen todas y cada una de las siguientes condiciones: • Tienen componentes semánticos compatibles. • Están asociados típicamente entre sí. • Son miembros correspondientes de cada par (A1, A2, A3,..). • Pertenecen a la misma categoría léxica. • Llenan la misma posición sintáctica en una construcción sintáctica. • Tienen la misma función semántica.
Cardinalidad:	Definida según sea el caso.
Estructura adyacente:	Definida según sea el caso.
Dominio:	Clases o categorías.
Rango:	Varía según la relación, aunque puede ser cualquiera de los elementos definidos como su dominio.
Propiedades:	Definidas según sea el caso.
Superclase:	Relación léxica.
Subclases:	*Actor*, *situación*, *instrumento*, *bienhechor*, *meta*, *percepción*, *sonido*, *degradación*, *intensificación* y *composición material*.

Relación:	**Relación léxica paradigmática.**
Definición:	Relación léxica entre unidades léxicas que comparten uno o más componentes semánticos de su núcleo, pertenecen a la misma categoría léxica, llenan la misma posición en una construcción sintáctica y tienen la misma función semántica.
Cardinalidad:	Definida en sus subrelaciones.
Estructura adyacente:	Definida en sus subrelaciones.

Dominio:	Definido en sus subrelaciones.
Rango:	Definido en sus subrelaciones.
Propiedades:	Definidas en sus subrelaciones.
Superclase:	Relación léxica.
Subclases:	*Sinonimia*, *oposición, hiperonimia y meronimia*.

Relación:	**Sinonimia.**
Definición:	Relación léxico paradigmática entre dos o más unidades léxicas, las cuales tienen el mismo núcleo de componentes semánticos y difieren sólo en sus componentes periféricos o superficiales.
Cardinalidad:	Muchos a muchos.
Estructura adyacente:	Conjunto.
Dominio:	Palabra.
Rango:	Un significante de la misma categoría que el dominio.
Propiedades:	Transitiva, reflexiva, conmutativa y su relación inversa es la oposición.
Superclase:	Relación léxico paradigmática.
Ejemplo(s):	Los sinónimos de un sustantivo son otros sustantivos (*enojo – enfado – irritación*). Los sinónimos de un adjetivo son otros adjetivos (*sabroso – apetitoso – exquisito*). Los sinónimos de un verbo son otros verbos (*requerir – pedir – solicitar*). Los sinónimos de un adverbio son otros adverbios (*también – además – asimismo*).

Relación:	**Oposición.**
Definición:	Relación léxica paradigmática entre dos o más synsets los cuales tienen un núcleo de componentes semánticos opuestos en diferentes contextos.
Cardinalidad:	Uno a uno.
Estructura adyacente:	Par.
Dominio:	Synsets adjetivo, synsets sustantivo, synsets verbo y synsets adverbio.
Rango:	Igual al dominio que se use.
Propiedades:	Relación inversa de la sinonimia. La propiedad de oposición se hace transitiva entre los sinónimos que conforman el synset del dominio con respecto a los sinónimos del synset definido en el rango.
Superclase:	Relación léxico paradigmática.
Subclases:	*Complemento*, *antonimia* y *convención direccional*.

Relación:	**Superordenación-subordinación, Hiperonimia-hiponimia, Genérico-específico.**
Definición:	Relación léxica paradigmática en la que uno de los synsets (denominado hiperónimo) no posee ningún rasgo semántico, o sema, que no comparta otro de los synsets (denominado hipónimo); mientras que éste sí posee rasgos semánticos que la diferencian de aquél.
Cardinalidad:	Uno a muchos.
Estructura adyacente:	Árbol.
Dominio:	Synset adjetivo, synset sustantivo, synset verbo, synset adverbio, y otros (hiperónimo).
Rango:	Igual al dominio que se use (hipónimo).

Propiedades:	Transitividad entre hipónimos, transitividad entre hiperónimos, reflexividad entre hiperónimos e hipónimos, rasgos semánticos comunes entre cohipónimos.
Superclase:	Relación léxico paradigmática.
Ejemplo(s):	*Coche1* posee sólo los semas [+vehículo], [+con motor] y [+pequeño tamaño], que comparte con *descapotable1*, mientras que *descapotable1* posee además el rasgo [+sin capota], que lo diferencia de *Coche1*.

Relación:	**Parte-todo o meronimia-holonimia.**
Definición:	Relación léxica paradigmática que se centra en las unidades léxicas usadas para describir partes y todos. Meronimia es la relación semántica entre unidades léxicas, o también synsets, que denotan partes y unidades léxicas o también synsets, que denotan el correspondiente todo. Todas las unidades léxicas que denotan partes de un todo son entre sí comerónimas.
Cardinalidad:	Uno a muchos.
Estructura adyacente:	Árbol.
Dominio:	Synset sustantivo (holónimo).
Rango:	Synset sustantivo (merónimo).
Propiedades:	Transitividad entre merónimos, transitividad entre holónimos y reflexividad entre merónimos-holónimos.
Superclase:	Relación léxico paradigmática.
Subclases:	*Componente-objeto integrado, miembro-colección, porción-masa, materia-sustancia, acción-actividad, lugar-área y topológica.*

ANEXO B: **Detalles de la prueba de afinación**

Objetivo.- Realizar una prueba exploratoria a la propuesta de utilizar ontologías para configurar una interfaz en lenguaje natural hacia bases de datos, para refinar ciertos aspectos que se consideran como críticos para la aceptación de la misma por parte de usuarios potenciales.

Esta prueba se diseñó en base a otra prueba similar, realizada en un proyecto doctoral [11], cuyo objetivo fue comparar la mejora que se tenía con una metodología de diseño de bases de datos distribuidas propuesta con respecto a la metodología considerada mejor hasta ese momento.

Un detalle importante es que los sujetos de prueba venían de un centro de investigación, por lo cual se les considera más preparados y mejor seleccionados que los alumnos de la universidad privada que hicieron la prueba No. 1, aunque con mucha menos experiencia laboral.

B.1 Descripción de la prueba de afinación

Se hizo una encuesta para definir el perfil de los elementos del grupo (tabla B-1) y se evaluaron tres aspectos: el entrenamiento recibido, la metodología propuesta y su experiencia al realizar la tarea de diseñar los cambios para que la ontología contenga la información necesaria para contestar todas las consultas del usuario. Cada uno de los tres aspectos anteriores fue evaluado a partir de una serie de preguntas evaluadas según la escala de Likert.

Tabla B-1 Datos generales.

Lugar	Cenidet.
Tipo de selección de alumnos	Proceso de selección riguroso.
Nivel de estudios	Maestría.
Periodo de pruebas	26-28/03/06.
Capacitación previa	Una explicación (45 minutos) definiendo el concepto de ontologías. Además, se proporcionó retroalimentación después de cada evaluación.
Material de apoyo	Manuales de las dos herramientas y documentación de la base de datos (se les entregó poco después de la explicación).

B.2 Resumen de la prueba de afinación

En la tabla B-2 se muestran las diferencias entre cada pregunta evaluada contra su evaluación anterior acerca del entrenamiento después de la prueba. Esto demuestra la mejoría obtenida en cada evaluación.

Tabla B-2 Diferencias en la evaluación del entrenamiento entre evaluaciones.

Pregunta	Evaluación No. 2 - Evaluación No. 1	Evaluación No. 3 - Evaluación No. 2
1) Considera que el entrenamiento para hacer la tarea de configuración fue el adecuado.	15.83	10.00
2) La sesión de entrenamiento le permitió entender la metodología de configuración de la interfaz.	3.33	16.67
3) Me sentí cómodo con el ambiente de trabajo después de la sesión de entrenamiento.	2.50	13.33
4) Fue sencillo de aprender a usar el editor Protégé.	-0.83	16.67
5) Considera que la terminología usada para configurar la interfaz es extraña o confusa.	12.50	16.67

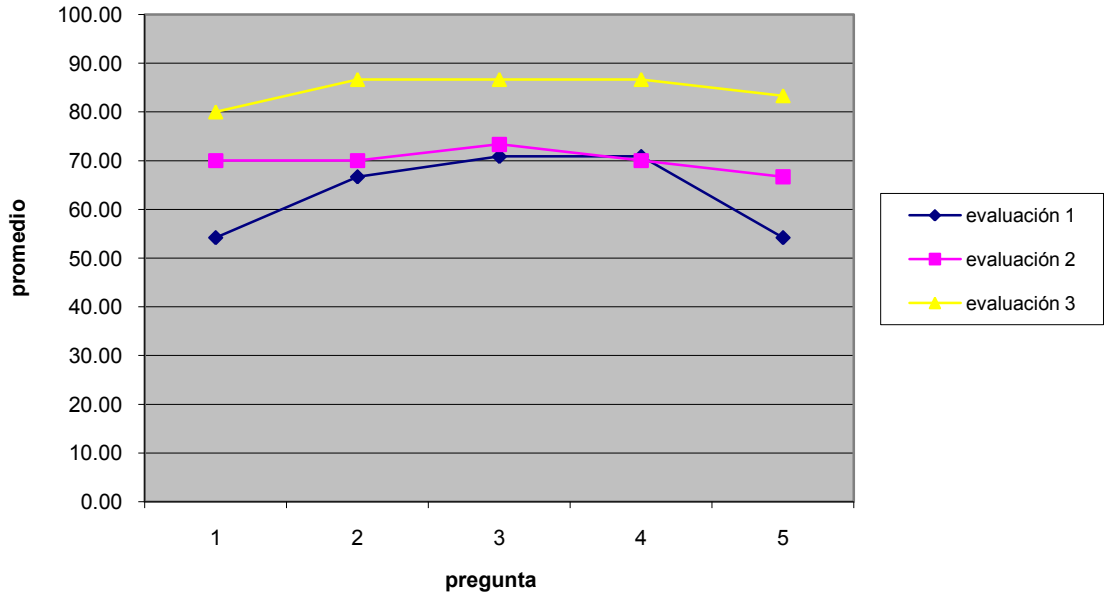

Figura B-1.- Gráfica de las tres evaluaciones del entrenamiento para la prueba de evaluación

En la tabla B-3 se muestran las diferencias entre cada pregunta evaluada contra su evaluación anterior acerca de la metodología. Esto demuestra la mejoría obtenida en cada evaluación

Tabla B-3 Diferencias en la evaluación del entrenamiento entre evaluaciones.

Pregunta	Evaluación No. 2 - Evaluación No. 1	Evaluación No. 3 - Evaluación No. 2
1) Es difícil imaginar las respuestas que contestará la interfaz.	-1.679	6.67
2) Entiendo la semántica de la base de datos.	10.83	3.33
3) La configuración fue compleja en términos de relacionar los conceptos a ser modelados.	-2.50	16.67
4) Estaban claros los pasos necesarios para llevar a cabo la configuración.	10.83	16.67
5) La configuración fue en un tiempo razonable.	-9.17	13.33
6) Se sintió cómodo al analizar la semántica de la base de datos.	-2.50	3.33
7) Se sintió cómodo al analizar y clasificar el corpus de preguntas.	1.67	13.33
8) Se sintió cómodo al analizar y completar los conceptos de la ontología.	3.33	26.67
9) Se sintió cómodo al analizar y completar las relaciones de la ontología.	0.83	26.67
10) El proceso de configuración es eficiente.	6.67	3.33
11) Considera que la configuración que realizó fue completa.	8.33	3.33

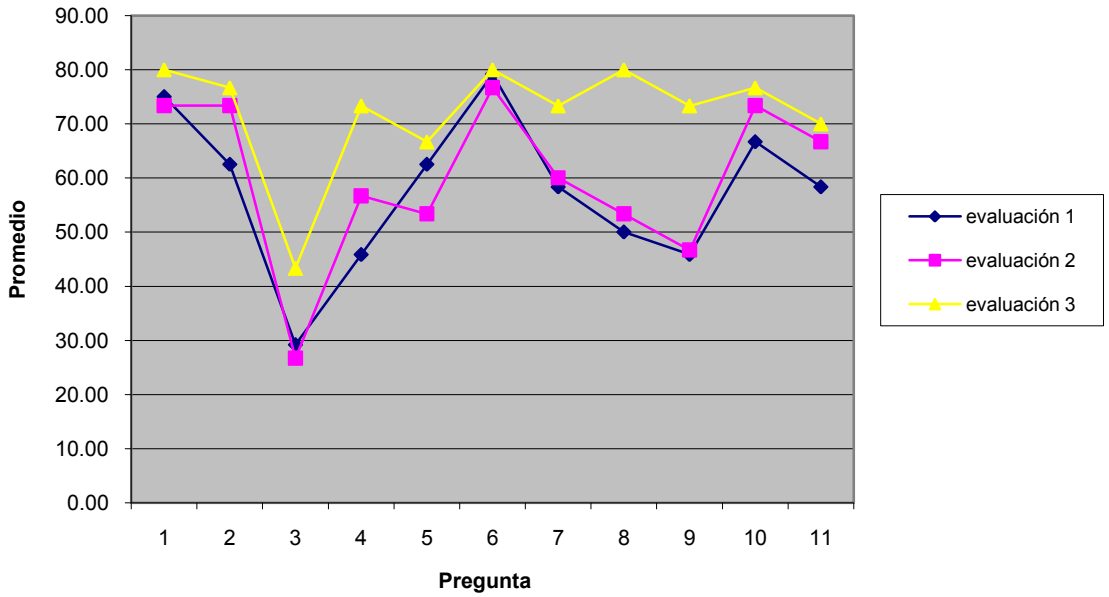

Figura B-2.- Gráfica de las tres evaluaciones de la metodología para la prueba de evaluación

En la tabla B-4 se muestran las diferencias entre cada pregunta evaluada contra su evaluación anterior acerca de la tarea de configuración. Esto demuestra la mejoría obtenida en cada evaluación.

Tabla B-4 Diferencias en la evaluación del entrenamiento entre evaluaciones.

Pregunta	Evaluación No. 2 - Evaluación No. 1	Evaluación No. 3 - Evaluación No. 2
1) Considera que la configuración es parecida a una aplicación real.	13.33	3.33
2) Considera que la configuración de la interfaz es flexible (susceptible de cambios o variaciones según las circunstancias o necesidades).	-0.83	13.33
3) Considera que la configuración de la interfaz es manejable (moverse con cierta soltura después de haber tenido algún impedimento).	-3.33	10.00
4) Considera que la configuración de la interfaz es inteligible (que puede ser entendida).	-3.33	16.67
5) Considera que la documentación de la configuración de la interfaz es fácil de entender (tener idea clara de las cosas).	-14.17	20.00

6) Considera que la configuración le da una idea de cómo funciona la interfaz (cómo ejecuta las funciones que le son propias).	-5.83	10.00
7) Afecta en la configuración el que en la interfaz haya elementos que no son en su idioma nativo.	20.00	16.67

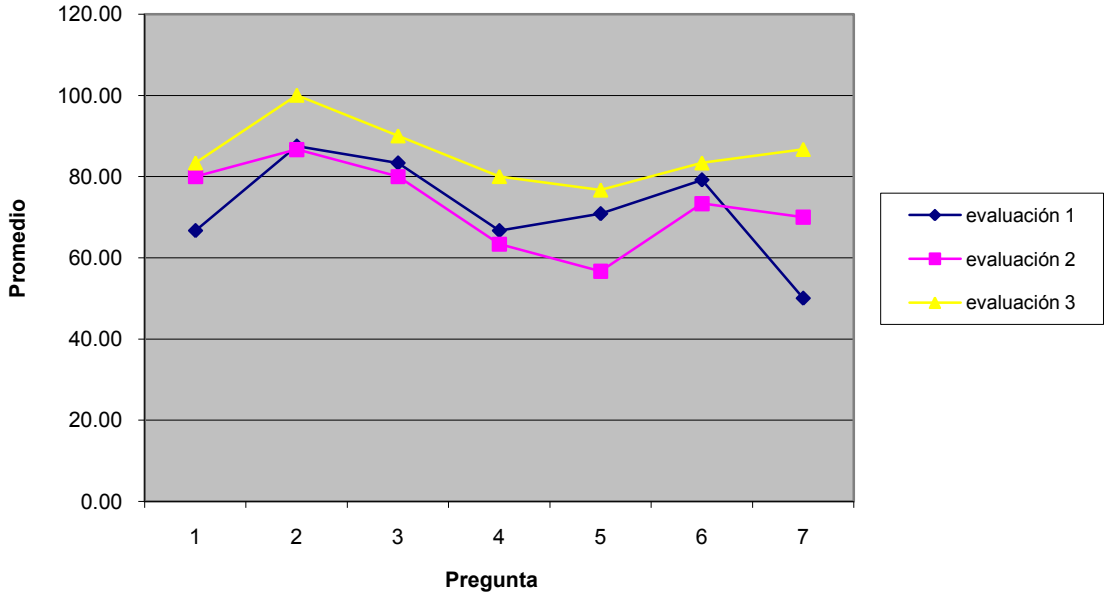

Figura B-3.- Gráfica de las tres evaluaciones de la configuración para la prueba de evaluación

Tabla B-5 Valores característicos.

Pregunta	Evaluación No. 1	Evaluación No. 2	Evaluación No. 3
Mejor calificación promedio	87.5 (Considera que la configuración de la interfaz es flexible)	86.67 (Considera que la configuración de la interfaz es flexible)	100 (Considera que la configuración de la interfaz es flexible)
Peor calificación promedio	29.17 (La configuración fue compleja en términos de relacionar los conceptos a ser modelados)	26.67 (La configuración fue compleja en términos de relacionar los conceptos a ser modelados)	43.33 (La configuración fue compleja en términos de relacionar los conceptos a ser modelados)

Diferencias entre el entrenamiento antes de la prueba y el de las evaluaciones	-6.37	0.29	14.96

Nota: Un valor positivo en las columnas evaluación No. 1, evaluación No. 2 o evaluación No. 3, significa que fue mejor la opinión que se tenía del entrenamiento después de realizar la prueba y un valor negativo significa lo contrario.

Tabla B-6 Resumen de las diferencias entre las evaluaciones 1, 2 y 3

Diferencia	Evaluación No. 2 - Evaluación No. 1	Evaluación No. 3 - Evaluación No. 2
La más alta	20 (Afecta en la configuración el que la interfaz haya elementos que no son en su idioma nativo)	26.67 (Se sintió cómodo al analizar y completar los conceptos y relaciones de la ontología)
La más baja	-0.83 (Considera que la configuración de la interfaz es flexible)	3.33 (5 factores)
Evaluación del entrenamiento	6.67	14.67
Evaluación de la metodología	2.42	12.12
Evaluación de la tarea de configuración	0.83	12.86

Análisis de la prueba de afinación

a) Las diferencias entre las evaluaciones a las preguntas mejoran en general en la tercera evaluación (Fig. B-1, Fig. B-2 y Fig. B-3), siendo lo más destacado la mejora de la evaluación de la tarea de configuración (Fig. B-3), donde un aumento de la complejidad permite sólo una mejora de 0.83 entre la primera y la segunda prueba, mientras que un retroceso de la complejidad entre la segunda y la tercera prueba permite una mejora de 12.86.

b) Las diferencias entre la primera y la segunda evaluación a las preguntas acerca del entrenamiento es únicamente negativa en el aprendizaje del editor (-0.83), debido a que en la segunda evaluación se necesitó un mayor uso y conocimiento del editor, por lo complejo de la tarea a realizar (Fig. B-1).

c) Las diferencias entre las evaluaciones a las preguntas acerca de la metodología mejoran en general en la tercera evaluación, aunque no de la misma manera, ya que la evaluación de lo adecuado del entrenamiento y los problemas generados por tener una interfaz con elementos en inglés avanzan en general entre diferentes evaluaciones, pero más significativamente entre la primera y segunda evaluación que entre la segunda y tercera evaluación (Fig. B-2).

d) Las diferencias entre la primera y la segunda evaluación a las preguntas acerca de la metodología son negativas en aspectos de complejidad (-2.5), tiempo invertido para configurar (-9.17) y la comprensión de la semántica de la base de datos (-2.5). Entre la segunda y la tercera evaluación sólo es negativa la diferencia relacionada con lo claro que es para el configurador relacionar su trabajo con los resultados de la ILNBD (-6.67), siendo también este aspecto el peor calificado en las primeras dos evaluaciones (Fig. B-2).

e) La flexibilidad fue el aspecto mejor evaluado, teniendo en la última evaluación una calificación unánime de 100.

f) En general las diferencias en la evaluación del entrenamiento, la metodología y la tarea de configuración mejoran notablemente entre la segunda y tercera evaluación en un promedio de 9.16, siendo la diferencia más significativa de 12.03, entre la evaluación de las diferencias de la tarea de configuración entre la primera y segunda evaluación (0.83) y entre la segunda y la tercera evaluación (12.86).

g) Desde la segunda evaluación, mejora la opinión que tenían los sujetos de prueba de la capacitación antes de empezar la prueba y después de realizarla, alcanzándose una mejora de 14.96 en la tercera evaluación (Fig. B-1).

Conclusiones de la prueba de afinación

- Las diferencias encontradas al punto *a* de la sección anterior, están asociadas a la curva de aprendizaje, y el crecimiento diferente que tiene la evaluación de la capacitación está relacionado con el poco entrenamiento previo que se dio, lo cual es avalado por los valores del punto b, de la sección anterior.
- De igual manera, las diferencias negativas encontradas entre la segunda y la primera evaluación están asociadas a la curva de aprendizaje y a que en la segunda evaluación la complejidad de las preguntas a configurar es mayor, lo cual se manifiesta en que la diferencia de la complejidad y del tiempo en que se llevó a cabo la configuración sean negativos. Este punto también es avalado por el análisis de la sección anterior en el punto d. La confusión que se dio en general con la terminología usada en la tercera evaluación está asociada al cambio de base de datos con respecto a las dos primeras evaluaciones.

- La sensación de que se completaron mejor las primeras dos configuraciones que la tercera está asociado al cambio de base de datos a consultar.
- Debido a que la propuesta no se configura en una ILNBD completa, es comprensible que no se imagine fácilmente el resultado de la configuración, a diferencia de English Query, que sí es una ILNBD completa.
- Es natural que en la tarea de configuración, a pesar de la curva de aprendizaje, hayan diferencias negativas, ya que además de que aún no se había asimilado bien la metodología de configuración basado en ontologías, se incrementó la complejidad en la tarea de configuración de la primera evaluación a la segunda, demostrándose que este factor influye bastante en la evaluación de la tarea de configuración al mejorar bastante entre la segunda y la tercera evaluación al regresar a la complejidad de las preguntas a configurar de la primera evaluación (Fig. B-3).
- La flexibilidad continúa siendo, igual que en la primera prueba, el aspecto en que más se aprecia la mejora de la propuesta de usar ontologías para configurar la ILNBD, llegándose en la tercera evaluación a la máxima calificación posible de forma unánime.
- Para esta prueba, la evaluación de la metodología, el entrenamiento y la tarea de configuración es dependiente de la experiencia que se tenga, la calidad de la documentación disponible, y de la complejidad de las consultas a configurar.

ANEXO C: **Ejemplo de la configuración de una consulta con la metodología propuesta**

Pregunta 1: Lista los clientes del ZIP “t2f”

Lista, junto con mostrar, enseñar y otros verbos, no aportan nada a la oración (fig. C-1).

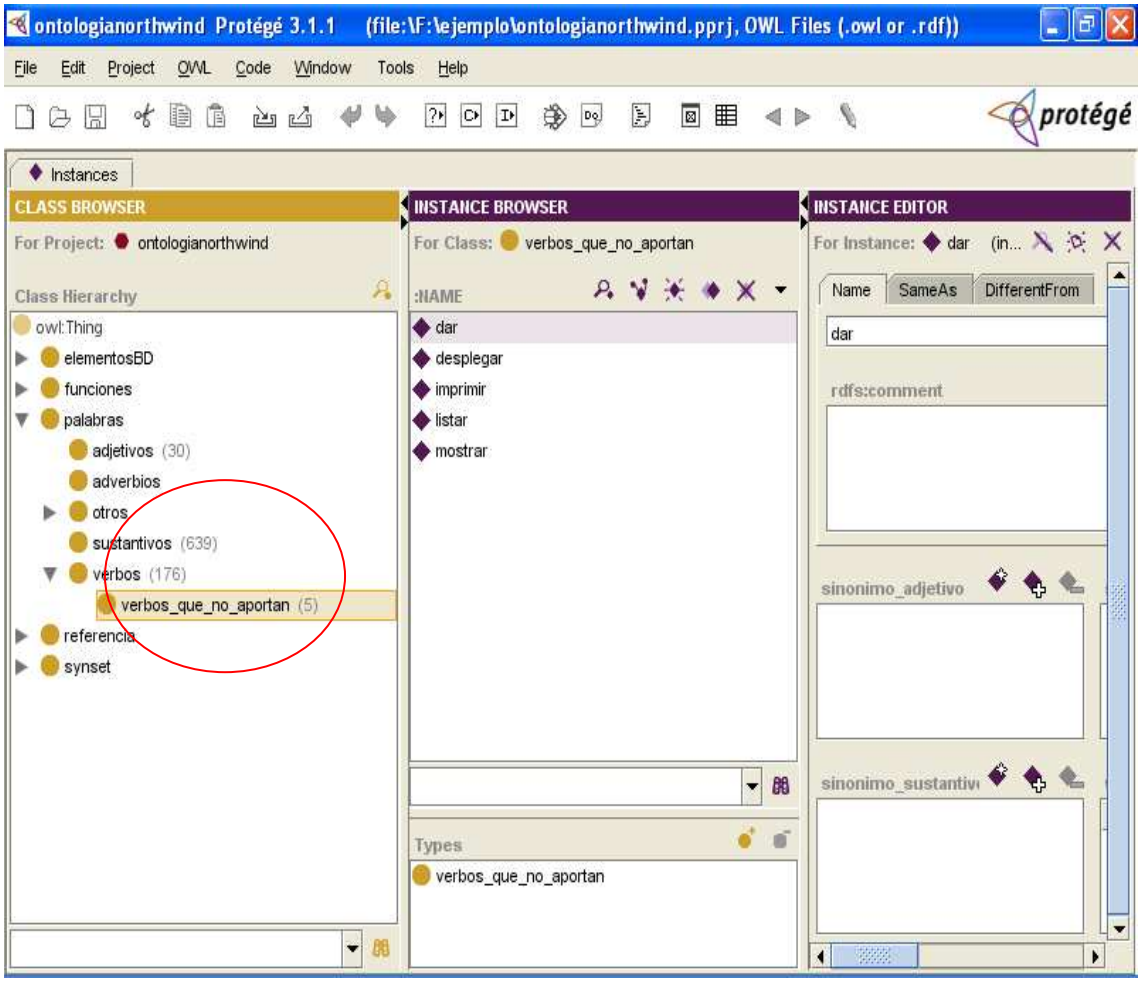

Figura C-1.- Verbos que no aportan

Revisar el ejemplar de la clase tabla clientes (Customers) y ver con qué synsets se relaciona (fig. C-2).

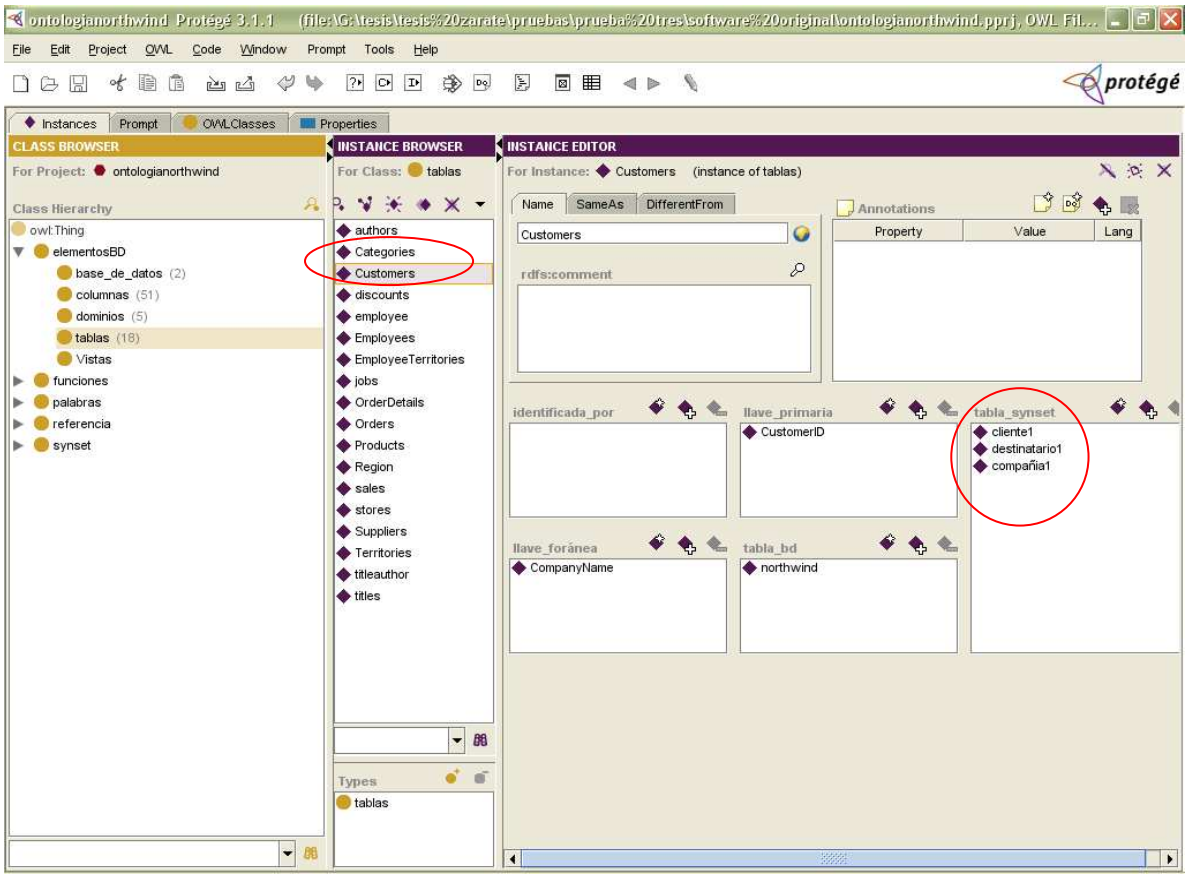

Figura C-2.- Relaciones del synset Customers

Dar doble clic en cliente1 y revisar los sustantivos con los que se relaciona (fig. C-3).

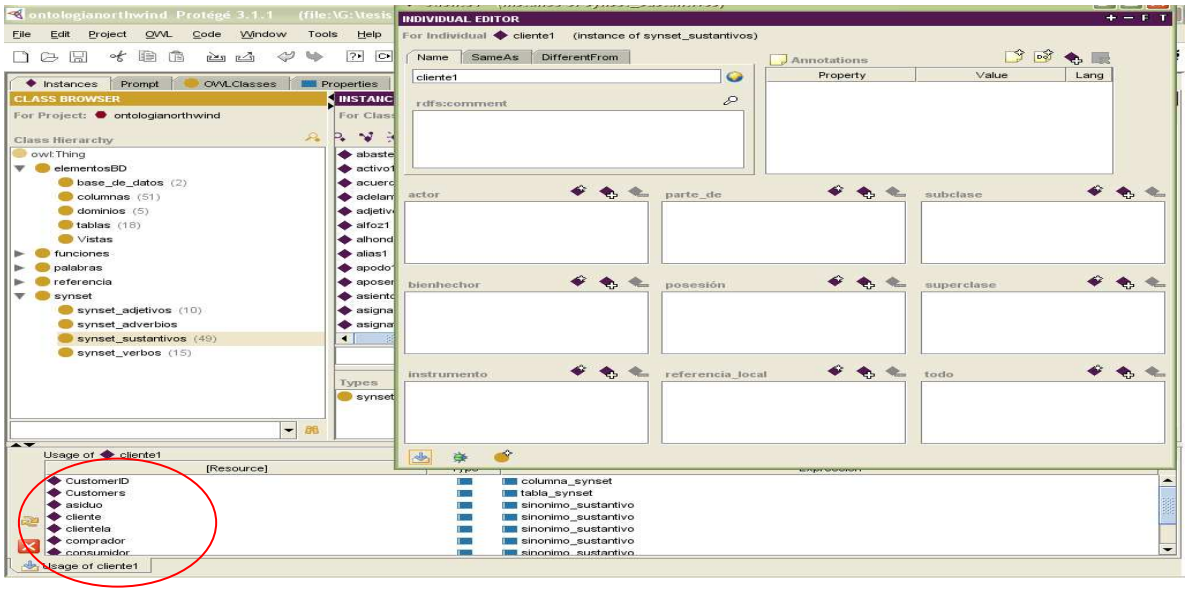

Figura C-3.- Relaciones del synset cliente1

Se encuentra que cliente está relacionado con cliente1 y éste a su vez con customer, y por transitividad customer se relaciona con cliente (fig. C-3).

Dar doble clic en la tabla clientes (customers) y revisar dónde se guarda su código postal (fig. C-4).

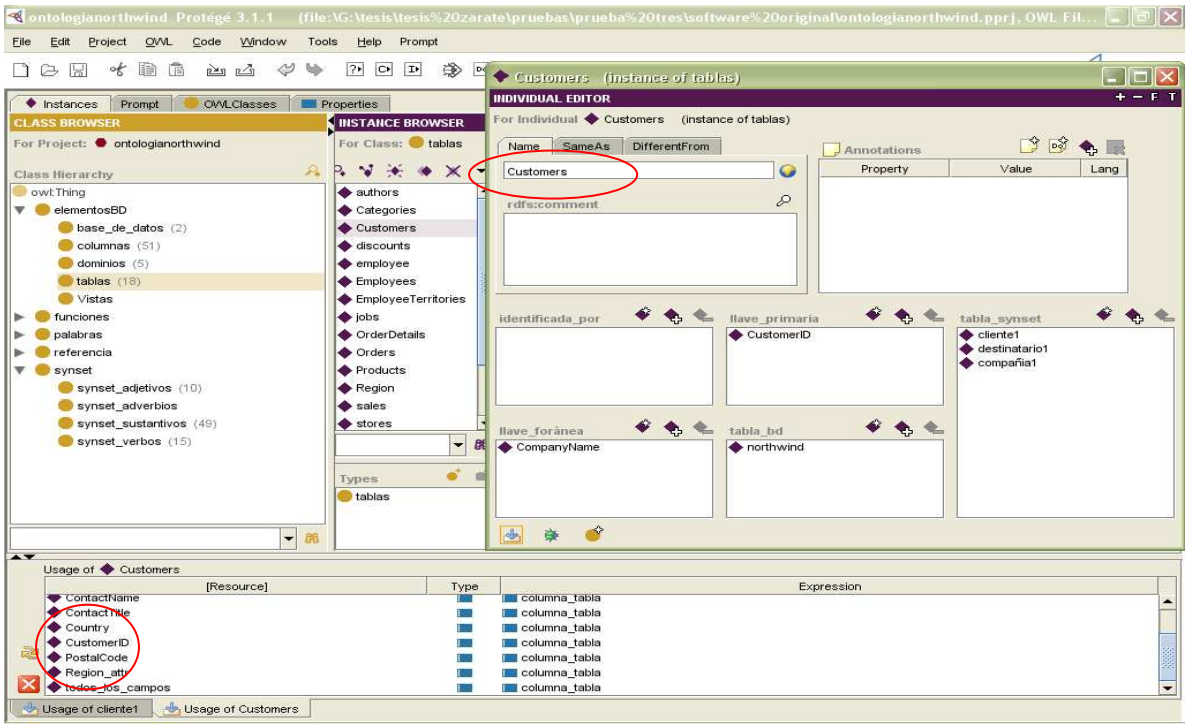

Figura C-4.- Código postal de Customers

Dar doble clic en postalcode y buscar los synsets con los que se relaciona (fig. C-5).

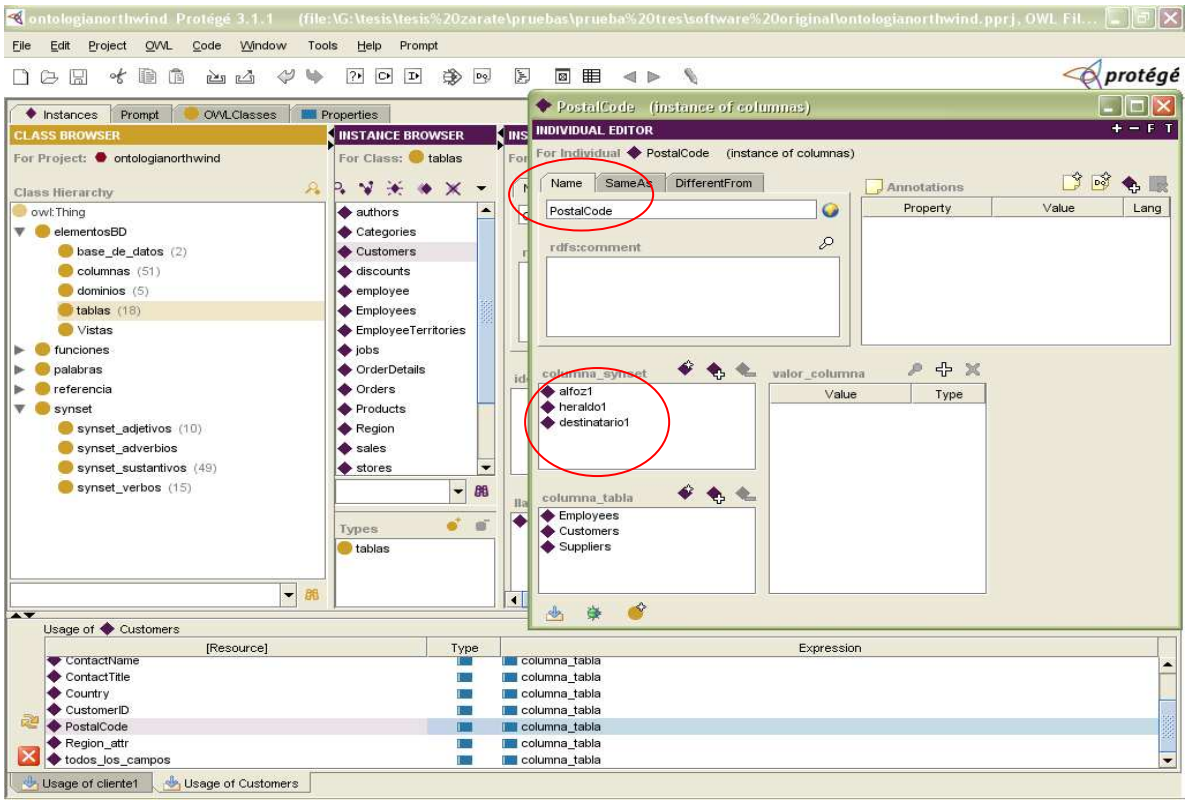

Figura C-5.- Relaciones del synset PostalCode

Dar doble clic en Heraldo1 y revisar los sustantivos con los que se relaciona (fig. C-6).

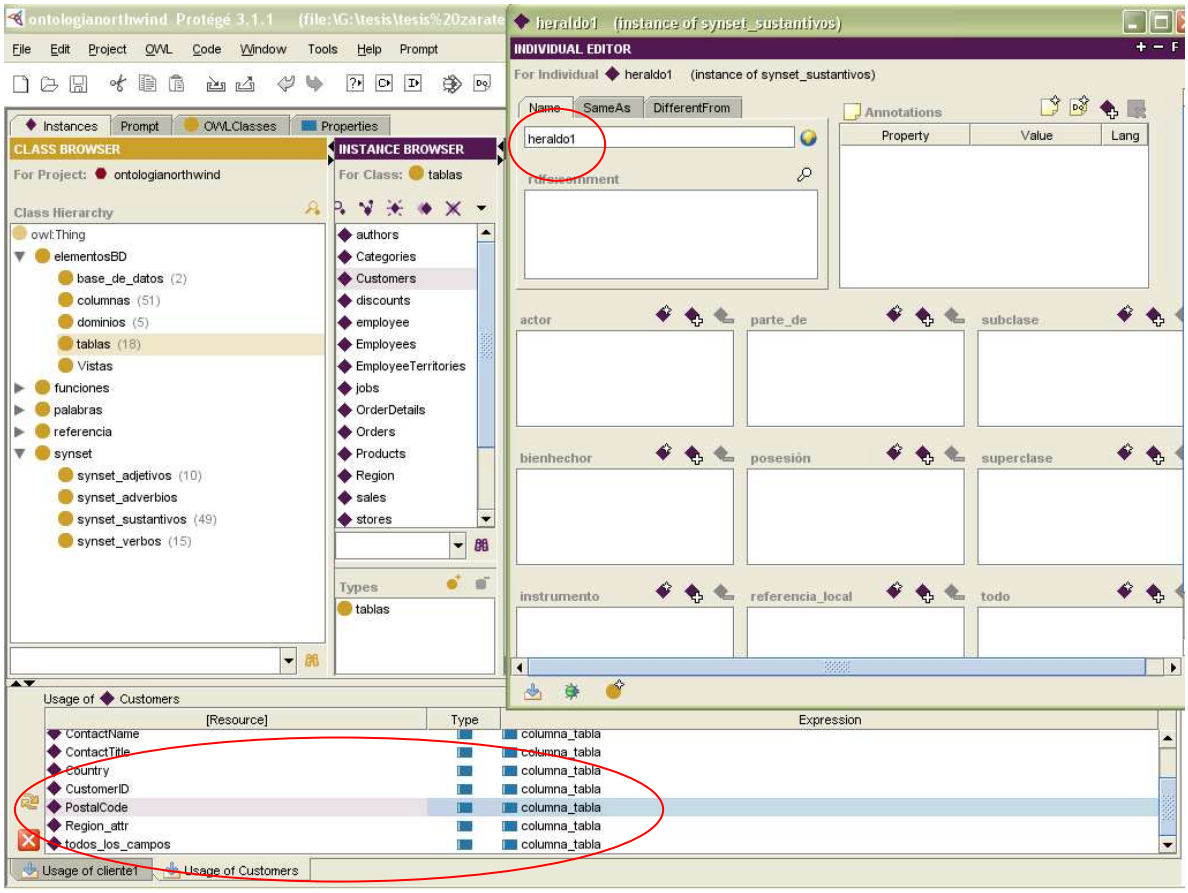

Figura C-6.- Relaciones del synset Heraldo1

Como zip no existe, se crea un ejemplar de sustantivo y se relaciona con heraldo1, el cual está relacionado con postalcode, y por transitividad se logra que zip y postalcode se relacionen (fig. C-7).

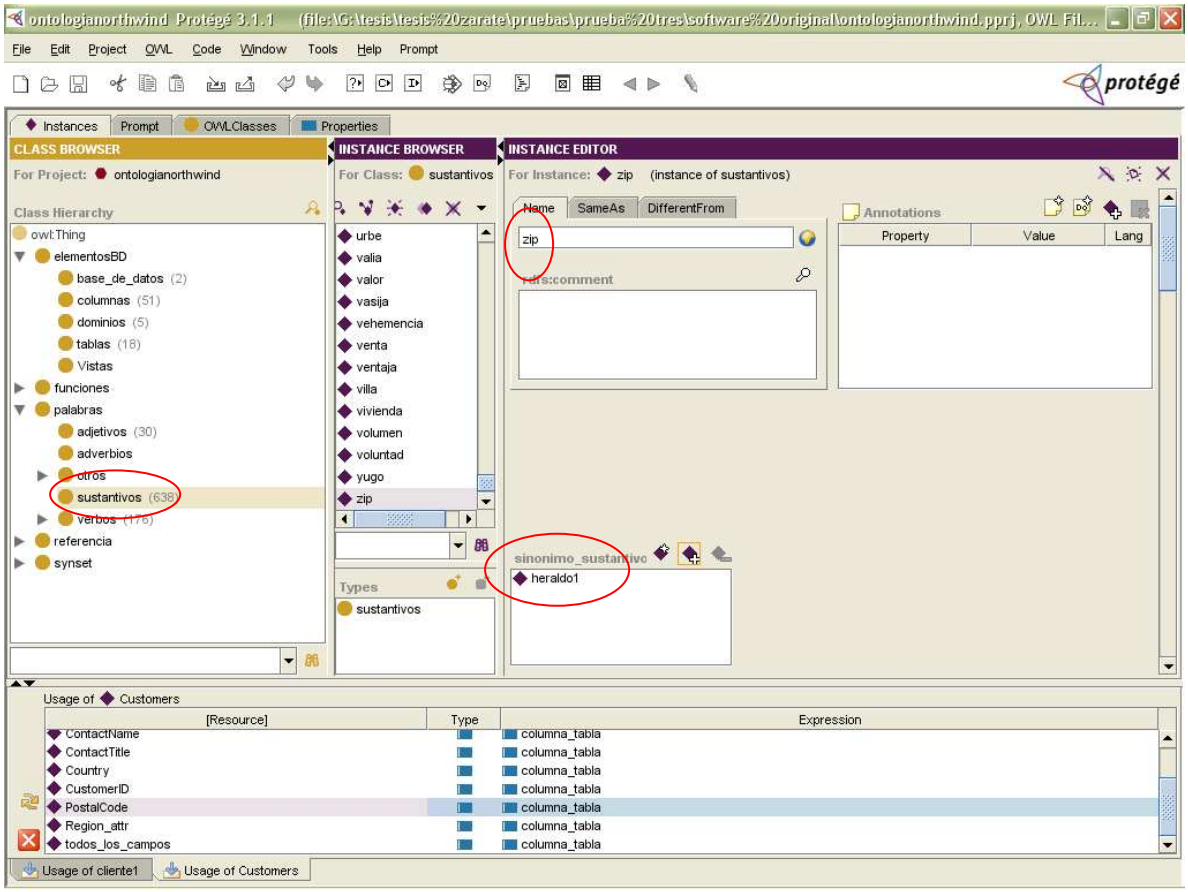

Figura C-7.- Creación del ejemplar de la clase sustantivos zip

ANEXO D: **Ejemplo de un cuestionario de obtención de perfil**

Objetivo: se busca conocer las características de los individuos que hacen la prueba.

Por favor lee las instrucciones cuidadosamente y responde todas las preguntas.

Tu cooperación es de gran ayuda para una investigación.

Datos generales

Edad: 23

Sexo: Masculino

Años de experiencia laboral: 4 años

Promedio acumulado de la maestría: 8.0

1.- Marque con una palomita los cursos que ha llevado en maestría o en licenciatura de la siguiente lista:

- Diseño de bases de datos. √
- Bases de datos I. √
- Análisis y diseño de sistemas. √
- Análisis orientado a objetos.
- Representación del conocimiento.
- Procesamiento de lenguaje natural.

2.- Por cuánto tiempo te has desempeñado y en qué papeles con respecto a un sistema de base de datos:

- Usuario: años.
- Analista/diseñador: años.
- Programador: 2 años. √
- Administrador: años.
- Otros: años , especifica_____

3.- Qué modelos de diseño de bases de datos conoces:

- Relacional. √
- Entidad – relación. √
- Orientado a objetos.
- Semántico.
- Semiestructurado con XML.

4.- Has diseñado alguna vez una base de datos:

– No.
– Sí, especifica en qué condiciones fue (trabajo de escuela, sistema real y qué metodología y herramientas usaste). Access, SQLServer, Oracle, Fox

5.- Autoevaluando tus habilidades y experiencia en el diseño de bases de datos, cómo te consideras:

1 2 3 4 5 6 7

Experto Novato

6.- Autoevaluando tus habilidades y experiencia en la teoría de bases de datos, cómo te consideras:

1 2 3 4 5 6 7

Conozco mucho Conozco poco

7.- Autoevaluando tus habilidades y experiencia en el análisis y diseño de sistemas, cómo te consideras:

1 2 3 4 5 6 7

Experto Novato

8.- Autoevaluando tu interés en el análisis y diseño de sistemas, cómo te consideras:

1 2 3 4 5 6 7

Muy interesado Poco interesado

9.- Cuando aprendes una nueva metodología de análisis y diseño, cómo te consideras:

1 2 3 4 5 6 7

Mente abierta Conservador

1 2 3 4 5 6 7

Muy entusiasta Poco entusiasta

1 2 3 4 5 6 7

Aprendo rápidamente Aprendo lentamente

10.- ¿Qué tanto usas en promedio la computadora al día?

1. Menos de una hora.
2. Entre 1-3.
3. Entre 4-6.
4. Entre 6-9.
5. Más de 9 hrs.

11.- Autoevaluando tus habilidades para aprender a usar una nueva herramienta, cómo te consideras:

1 2 3 4 5 6 7

Aprendo rápidamente Aprendo lentamente

ANEXO E: **Funciones desarrolladas para explotar la ontología**

Class rsesame3

En esta clase del lenguaje Java se definen los métodos para acceder a la ontología genérica, a través de una extensión a la clase Sesame.

Public Class rsesame3
Extends sesame

Ver también:

Resumen de campos	

Resumen del constructor
InitSesame() Define un objeto del repositorio Sesame local, con los valores por omisión.

Resumen de los métodos	
String	consParteDe(String query1) Regresa una cadena de los sustantivos que son parte de *query1.*
String	conssinonimo(String query1, int pos) Regresa una cadena de los sinónimos de *query1.*
String	conshiponimo(String query) Regresa una cadena de los hipónimos de *query.*
Vector	conshiperonimo(String query) Regresa una cadena de los hiperónimos de *query.*
Vector	consMiembroDe(String query) Regresa una cadena de los sustantivos que son miembros del grupo

	definido por *query.*
Vector	consHechoDe(String query) Regresa una cadena de los sustantivos que representan los materiales de los que está hecho *query.*
Vector	consSesame(String query) Regresa una cadena de los sustantivos que son parte de *query.*
Vector	consantonimo(String query, int pos) Regresa una cadena de los antónimos de *query.*

Métodos heredados de la clase sesame:

Detalle del constructor

public String **InitSesame**()

Define un objeto del repositorio Sesame local, con los valores por omisión.

Detalle de los métodos

public String consParteDe(String query1)

Regresa una cadena de los sustantivos que son parte de *query1.* Sólo los sustantivos presentan esta propiedad.

Parámetros:

query1.

Regresa:

Una cadena de sustantivos separados por comas.

Ver también:

Rsesame3.consMiembroDe, Rsesame3.consHechode.

public String conssinonimo(String query1, int pos)

Regresa una cadena de sinónimos de la palabra definida en *query1*. Los sinónimos son palabras que tienen el mismo significado, aunque se escriban de diferente manera. Los sinónimos pertenecen a la categoría definida por *pos*, es decir si *pos* es 2 regresará los verbos sinónimos de *query1.*

Parámetros:

query1: palabra de la cual se quiere encontrar sus sinónimos.

pos: identificador de la parte del habla: 1) sustantivo, 2) verbo, 3) adjetivo, 4) adverbio.

Regresa:

Una cadena de palabras separadas por comas.

Ver también:

Rsesame3.conshiponimo, Rsesame3.conshiperonimo, Rsesame3.consantonimo.

public static String consantonimo(String query, int pos)

Regresa una cadena de antónimos de la palabra definida en *query*. Los antónimos son palabras cuyo significado son contrarios y pertenecen a la categoría definida por el parámetro *pos*, es decir si *pos* es 2 regresará los verbos antónimos de *query.*

Parámetros:

query: palabra de la cual se quiere encontrar sus antónimos.

pos: identificador de la parte del habla: 1) sustantivo, 2) verbo, 3) adjetivo, 4) adverbio.

Regresa:

Una cadena de palabras separadas por comas.

Ver también:

Rsesame3.conshiponimo, Rsesame3.conshiperonimo, Rsesame3.conssinonimo.

public static Vector conshiponimo(String query)

Regresa una cadena de hipónimos de la palabra definida en *query*. Los hipónimos son palabras cuyo significado es una especialización del significado de *query*. Sólo los sustantivos presentan esta propiedad.

Parámetros:

query: palabra de la cual se quiere encontrar sus hipónimos.

.

Regresa:

Una cadena de hipónimos separados por comas.

Ver también:

Rsesame3.conssinonimo, Rsesame3.conshiperonimo.

public static Vector conshiperonimo(String query)

Regresa una cadena de hiperónimos de la palabra definida en *query*. Los hiperónimos son palabras cuyo significado es una generalización del significado de *query*. Sólo los sustantivos presentan esta propiedad.

Parámetros:

query: Palabra de la cual se quiere encontrar sus hiperónimos.

.

Regresa:

Una cadena de hiperónimos separados por comas.

Ver también:

Rsesame3.conshiponimo, Rsesame3.conssinonimo.

public static Vector consMiembroDe(String query)

Regresa una cadena de los sustantivos que son miembros del grupo definido por *query.* Sólo los sustantivos presentan esta propiedad.

Parámetros:

query

Regresa:

Una cadena de sustantivos separados por comas.

Ver también:

Rsesame3.consParteDe, Rsesame3.consHechoDe.

public static Vector consHechoDe(String query)

Regresa una cadena de los sustantivos que representan materiales de los que está hecho *query.* Sólo los sustantivos presentan esta propiedad.

Parámetros:

query

Regresa:

Una cadena de sustantivos separados por comas.

Ver También:

Rsesame3.consMiembroDe, Rsesame3.consPartede.

Referencias

[1] I. Androutsopoulos, G. Ritchie. P. Thanisch, “Natural Language Interfaces to Databases, an Introduction“, http://citeseer.nj.nec.com/1natural.html, Mar. 2001

[2] I. Androutsopoulos, G. Ritchie, P. Thanisch: “An Efficient and Portable Natural Language Query Interfaz for Relational Databases”. *Memorias de la 6a. Conferencia en Aplicaciones de Inteligencia Artificial y Sistemas Expertos en Ingeniería y en la Industria*., Edimburgo, págs 327--330. Gordon and Breach Publishers Inc., Langhorne, PA, U.S.A, Jun. 1993. ISBN 2--88124--604--4.

[3] J. Burger. et al. Issues, “Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST 2001”. http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/qa_roadmap.pdf, Oct. 2005.

[4] E. F. Codd, “A Relational Model for Large Shared Data Banks”. *Communications of the ACM*, Vol. 13, num. 6: págs. 377--387, 1970.

[5] E. Chay C., *Una Interfaz en Lenguaje Natural en Español para Consultas a Bases de Datos*. Tesis de maestría ITESM, Campus Morelos, jun. 1990.

[6] DARPA's Information Exploitation Office, “Darpa Agent Markup Language Home Page”, http://www.daml.org/, Oct. 2007.

[7] ElfSoft, "English Language Front-End Software", http://www.elf-software.com, Oct. 2007

[8] Elfsoft, "ELF vs. English Query vs. English Wizard", http://www.elf-software.com/FaceOff.htm, Oct. 2007

[9] M. García Ch., J. A. Zárate M., R. A. Pazos R., “Editor para Generar Ontologías a partir de un Diccionario de Datos”, Memorias del *10mo. Congreso Internacional de Investigación en Ciencias Computacionales*, Oaxtepec, México, págs. 47-55, Oct. 2003.

[10] M. García de Quesada. “Ontologías: Representaciones Independientes de la Lengua”. http://elies.rediris.es/elies14/cap432.htm, Oct. 2007.

[11] Ch. Garcia H., *A Semantic-Based Methodology for Integrated Distributed Database Design:Toward Combined Logical and Fragmentation Design and Design Automation,* P.H. Dissertation, The University of Arizona, Arizona, Abr. 1992.

[12] J. J. González B., *Interfaz en Lenguaje Natural Independiente del Dominio Capaz de Procesar Dominios Complejos*, tesis de doctorado, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Mor., dic. 2005.

[13] T. Gruber, “A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, págs. 199-220, Ene. 1993.

[14] C. D. Hafner y K. Godden, “Portability of Syntax and Semantics in Datalog”. *ACM Transactions on Office Information System*s, vol. 3 num.2: págs. 141-164, abr. 1985.

[15] G. Hendrix, E. Sacerdoti, D. Sagalowicz, and J. Slocum. “Developing a Natural Language Interface to Complex Data”, *ACM Transactions on Database System*, Vol. 3, No. 2, págs. 105–147, Jun.1978.

[16] V. Octavio H., *Un Método para el Reconocimiento a Bases de Datos en Interrogaciones en Lenguaje Natural*, Tesis de Maestría ITESM, Campus Cuernavaca. Jun. 1989.

[17] E. E. Loos (general editor), S. Anderson (editor), D. H. Day Jr. (editor), P. C. Jordan (editor), and J. D. Wingate (editor), “Glossary of Linguistic Terms”, http://www.sil.org/linguistics/ GlossaryOfLinguisticTerms/, Jun. 2007.

[18] H. G. Lugo, “Introducción al Procesamiento del Lenguaje Natural”, http:\\www.mia.uv.mx\ ~hlugo\materias.html, Jun. 2005.

[19] B. Magnini , S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo and M. Rijke. *The Multiple Language Question Answering,* Reporte del CLEF 2003 Workshop, Ed. Springer-Verlag, 2003.

[20] Microsoft Corp., "Chapter 32 - English Query Best Practices", http://www.microsoft.com/ technet/prodtechnol/sql/2000/reskit/part9/c3261.mspx, Jun 2007

[21] G. Miller, "Wordnet, a Lexical Database, Cognitive Science Laboratory", http://www.cogsci .princeton.edu/~wn/, Jun. 2007.

[22] E. Morales: “Curso de Representación de Conocimiento”, http://w3.mor.itesm.mx/~rdec/node1.html, dic. 1999

[23] N. F. Noy, D. L. McGuinness, ”Desarrollo de Ontologias-101: Guía para Crear tu Primera Ontología”, traducido del inglés por: E. Antezana,, http://protege.stanford.edu/publications/ ontology_development/ontology101-es.pdf, Sept 2005.

[24] Ontology Interchange Language (OIL),”Welcome to OIL”, http://www.ontoknowledge.org/oil/, Oct. 2007.

[25] I. R. Ponce M. “Buscador de Documentos Web Basado en una Ontología”, Tesis de maestría, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Mor., dic. 2006.

[26] Precise, “The PRECISE Natural Language Interface to Databases”, http://cognews.com /1062409630/index_html. Oct. 2007.

[27] Diccionario en línea de la Real Academia Española, ”Real Academia Española", http://www.rae.es, Oct. 2007.

[28] Resource Description Framework, ”Resource Description Framework (RDF) Model and Syntax Specification”. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, Oct. 2007.

[29] B. Richa A, *Natural Language Interfaces: Comparing English Language Front End and English Query*, Master of Science thesis, Virginia Commonwealth University, Richmond, Virginia, Dic., 2004.

[30] G. R. Rocher S., *Traducción de Queries en Prolog a SQL*, Tesis de Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas-Puebla. Sept. 1999.

[31] J. Rojas, J. Torres, “A Survey in Natural Language Databases Interfaces”, Memoria del *8mo. Congreso Internacional de Investigación en Ciencias Computacionales*, Inst. Tecnológico de Colima, Colima, México, Nov. 2001, págs. 63-70.

[32] Russian Institute of Artificial Intelligence, "Inbase", http:// www.inbase.artint.ru/english/default-eng.asp, Jun. 2007

[33] V. Sethi: *Natural Language Interfaces to Databases: MIS Impact, and a Survey of Their Use and Importance,* Reporte de la Business Graduate School, University of Pittsburgh. Pittsburgh, PA 15260, abr. 1990.

[34] S. Sharoff, “SNOOP a System for Development of Linguistic Processors”, proc. of *the East-West Conference on Artificial Intelligence: From Theory to Practice (EWAIC93),* Moscow, págs. 184-188, Sept. 1993.

[35] Stanford Medical Informatics, Stanford University, "Protégé Ontology Editor", http://protege.stanford.edu/ index.html, Jun. 2007.

[36] J. L.Vicedo, H. Rodríguez, A. Peña. y M. Massot. “Los Sistemas de Búsqueda de Respuestas desde una Perspectiva Actual”. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, n.31, 2003.

[37] Web Ontology Language, “Web Ontology Language (OWL),”, http://www.w3.org/2004/OWL/. Oct. 2007.

[38] W. A. Woods, J. Schmolze, “The KLONE family” in *Computers and mathematics with applications*, Vol.23: págs. 2-5, 1993.

[39] J. A. Zárate M., R. A. Pazos R., A. Gelbukh K., J. Pérez O.,”Improving the Customization of Natural Language Interface to Databases Using an Ontology”, *Proc. International Conference on Computational Science and Its Applications (ICCSA 2007)*, Kuala Lumpur, págs. 424-435, Ago. 2007.

[40] J. A. Zarate M., R. A. Pazos R., R. Toledo, “Aquisition of Lexical-sintactic Relatioships from a Dictionary”, *11th International Congress on Computer Science Research*, Tlanepantla, Mexico, págs. 45-52, Sept. 2004.