Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: *OntoTag*

Guadalupe Aguado de Cea[§], Inmaculada Álvarez de Mon y Rego[‡], Antonio Pareja Lora[†]

[§]DLACT. Facultad de Informática, UPM.
Campus de Montegancedo, s/n. 28660-Madrid, España
lupe@fi.upm.es

[‡]DLACT. Escuela Universitaria de Ingeniería Técnica de Telecomunicación. UPM.
Carretera de Valencia, Km. 7 28031-Madrid, España
ialvarez@euitt.upm.es

[†]DSIP. Facultad de Informática. UCM
Ciudad Universitaria, s/n. 28040-Madrid, España
apareja@sip.ucm.es

[§,‡,†]Grupo de Procesamiento de Lenguaje Natural (LIA-PLN)
Laboratorio de Inteligencia Artificial (LIA) Facultad de Informática. UPM
Campus de Montegancedo, s/n. 28660-Madrid, España

Resumen

A instancias de lo que se ha dado en llamar la Web Semántica, la Inteligencia Artificial ha investigado exhaustivamente la anotación semántica de páginas web. La anotación (semántica) de textos se desarrolló primeramente en la Lingüística de Corpus; sin embargo, la Inteligencia Artificial, al centrarse en una anotación basada en ontologías, parece haber pasado por alto sus resultados. Este artículo muestra nuestras primeras experiencias en la integración de ambos campos, según las cuales una anotación híbrida (lingüística y ontológica) no sólo sería posible, sino también de gran utilidad, para hacer más comprensibles a un ordenador los documentos de la Web Semántica. Nuestro equipo de investigación está desarrollando *OntoTag*, un modelo de anotación multi-nivel (en principio, también multilingüe y de propósito general) basado en los estándares EAGLES y en la Semántica Ontológica, e implementado en lenguajes de marcado de última generación (RDF(S)/XML).

Palabras clave: Web Semántica, anotación, semántica, ontologías, Lingüística de Corpus, OntoTag.

1 Introducción.

Hoy en día, todos estamos acostumbrados a usar de manera exhaustiva la llamada *World Wide Web* (WWW). La WWW ha sido y es una gran fuente de información accesible mediante nuestros ordenadores pero, hasta ahora, sólo comprensible para los seres humanos. Al principio, los documentos de la WWW se hacían a mano y nacían orientados al intercambio de información entre las personas. Todos estos documentos contenían una enorme cantidad de texto, imágenes e incluso sonido, sin significado para una computadora. El usuario de la WWW era el encargado de extraer e interpretar la información relevante. Actualmente, dado el asombroso crecimiento de la información contenida en Internet resulta imposible que un único usuario realice esas tareas en un tiempo aceptable. Al mismo tiempo, han surgido nuevas tecnologías que facilitan la gestión y recuperación de la información y, con ellas, aparece también la generación semi-automática de documentos web.

En la actualidad, la presentación en la WWW de este tipo de documentos tiende a tratarse independientemente de su contenido, principalmente mediante la utilización de XML (Bray *et al.*, 1998) u otros lenguajes orientados a los recursos como OML (Kent, 1998), XOL (Karp *et al.*, 1999), RDF (Lassila *et al.*, 1999), RDF Schema (Brickley *et al.*, 2000), SHOE (Luke *et al.*, 2000), OIL (Horrocks *et al.*, 2000) o DAML+OIL (Horrocks *et al.*, 2001). Pero aunque se está facilitando el procesamiento automático de la información, tampoco un ordenador puede llevar a cabo por sí solo las tareas ya mencionadas de acceso, extracción e interpretación de la información relevante. Por tanto, conseguir que las computadoras entiendan el significado (la semántica) de los textos escritos y de las páginas web es el pilar principal que sustenta el desarrollo de la *Web Semántica* (Berners-Lee *et al.*, 1999). En este contexto, la *anotación semántica de textos*, que hace explícito el significado para un ordenador, se ha convertido en un punto clave. Por esta razón, con el fin de conseguir extraer e interpretar la información de la mejor manera posible, se hace necesario para la anotación semántica de páginas web, tanto un diseño avanzado como la aplicación de modelos y formalismos. Últimamente, los especialistas en el área de desarrollo de ontologías han estudiado a fondo la anotación semántica de documentos web (Benjamins *et al.*, 1999; Motta *et al.*, 1999; Luke *et al.*, 2000; Staab *et al.*, 2000). Sin embargo, sus trabajos no han llegado a tener en cuenta, de alguna forma, los resultados obtenidos en la anotación de corpus en el campo de la *Lingüística de Corpus*, no sólo en el nivel semántico, sino también en otros niveles lingüísticos. Estos otros niveles lingüísticos, aunque no son intrínsecamente semánticos, contienen información adicional que puede ayudar a la computadora a entender un texto o, en nuestro caso, un documento de la WWW (Aguado *et al.*, 2002a; 2002b; 2002c; Buitelaar *et al.*, 2003; Martínez-Fernández y García-Serrano, 2002).

El objetivo del presente artículo es presentar los resultados de nuestra investigación sobre cómo la anotación lingüística puede ayudar a las computadoras a comprender la información textual en un documento de la Web Semántica. Se ha dedicado un esfuerzo especial a encontrar un modo de conjugar e identificar complementariedades entre los modelos de anotación semántica de la Inteligencia Artificial (IA) y las anotaciones propuestas por la Lingüística de Corpus (LC).

Este artículo está organizado como sigue: primero, en la sección 2, se presenta, a grandes rasgos, el uso actual de las ontologías en la anotación semántica. A continuación, se introduce el estado de la cuestión en la anotación de textos dentro de la LC (sección 3). En la subsección 3.1, se incluyen recomendaciones de alto nivel para la anotación lingüística , junto con una presentación de los principales niveles que dentro de ésta se consideran (subsección 3.2), a saber: anotación de lemas, morfosintáctica, sintáctica, semántica y de discurso. En la subsección 3.3 se enuncian las recomendaciones EAGLES sobre anotación sintáctica y morfosintáctica. En la sección 4, se presenta un ejemplo de integración de ambos paradigmas (IA y LC). Posteriormente se analizan las principales ventajas de esta integración (sección 5) y, finalmente, se enuncian algunas conclusiones y trabajos futuros pendientes (sección 6).

2 Anotaciones ontológicas y Web Semántica.

Los investigadores de la IA han encontrado en las *ontologías* (Gruber, 1993; Studer *et al.*, 1998) el modelo del conocimiento ideal para describir formalmente los recursos web y su vocabulario y, por tanto, para hacer explícito de algún modo el significado subyacente de los términos incluidos en las páginas web.

La *Semántica Ontológica* (Niremburg y Raskin, 2001) es una teoría que estudia el significado del lenguaje humano o lenguaje natural, así como una aproximación al Procesamiento del Lenguaje Natural (PLN) que utiliza un modelo abstracto del mundo –la ontología– como recurso central para extraer y representar el significado de textos en lenguaje natural, al razonar con el conocimiento que se deriva a partir de estos textos. Asimismo, la ontología es también el eje central a la hora de generar textos en lenguaje natural basados en las representaciones de su significado.

Con la Semántica Ontológica como punto de partida para la anotación de recursos web con información ontológica, se pretende conseguir que esta anotación permita el acceso inteligente a dichos recursos, facilite la búsqueda y navegación en la web y explote nuevos enfoques de inferencia a partir de estos recursos. Se han desarrollado muchos sistemas y proyectos en este sentido: la iniciativa $(KA)^2$ (Benjamins *et al.*, 1999); PlanetOnto (Motta *et al.*, 1999); SHOE (Luke *et al.*, 2000) y el proyecto *Semantic Community Web Portals* (Staab *et al.*, 2000). También han visto la luz hasta la fecha varias herramientas de anotación semántica: COHSE (COHSE, 2002), MnM (Vargas-Vera *et al.*, 2001), OntoMat-Annotizer (OntoMat, 2002), SHOE Knowledge Annotator (SHOE, 2002) y AeroDAML (AeroDAML, 2002). Un estudio del arte más extenso y general acerca de herramientas ontológicas se encuentra en OntoWeb (2002).

3 Anotación en Lingüística de Corpus.

La idea de *anotación* fue desarrollada originalmente en la Lingüística de Corpus. La **Lingüística de Corpus** (McEnery y Wilson, 2001) puede considerarse no una rama de la Lingüística en sí misma, como la sintaxis o la semántica, sino más bien una metodología o un enfoque, que pueden seguir estas otras ramas para explicar o describir un aspecto concreto de los usos del lenguaje. Según los mismos autores, la Lingüística de Corpus se aplicó por primera vez en la investigación sobre adquisición del lenguaje, para enseñar una segunda lengua o para elaborar gramáticas descriptivas, entre otras aplicaciones. Tradicionalmente, los lingüistas han definido *corpus* como "un conjunto de datos (auténticos) del lenguaje manifestados de forma natural y utilizables como base para la investigación lingüística" (Leech, 1997a). En la actualidad, el término **corpus** se aplica a "un conjunto de material lingüístico que existe en forma electrónica y que puede ser procesado por una computadora con distintos fines como la investigación lingüística y la ingeniería del lenguaje" (Leech, 1997a). Sin embargo, para que estos datos resulten útiles computacionalmente, han de estar anotados. Un **corpus anotado** "puede considerarse un almacén de información lingüística […] hecha explícita mediante una anotación concreta" (McEnery y Wilson, 2001). Las ventajas de una anotación de ese tipo está claro: hace más fácil y rápida la recuperación y el análisis de la información contenida en el corpus. Veamos ahora las recomendaciones en la Lingüística de Corpus para la anotación y los distintos niveles a los que es aplicable.

3.1 Recomendaciones generales para la anotación de textos.

En Leech (1997a) y McEnery y Wilson (2001) se presenta un conjunto de pautas, estándares o recomendaciones de buena práctica, aplicables a la anotación de textos (Ilustración 1). Por otra parte, los gobiernos y organismos internacionales han concedido financiación a diversos proyectos con miras a la unificación y estandarización de esquemas de anotación; un ejemplo de ello sería la iniciativa EAGLES de la UE (EAGLES 1996a; 1996b; 1996c). Uno de los resultados de esta iniciativa es el Estándar de Codificación de Corpus –*Corpus Encoding Standard* o CES (1999)– que incluye algunos criterios generales que deben considerarse cuando se elabora un esquema de anotación (Ilustración 1).

Un análisis detallado de todos estos criterios y conceptos puede encontrarse en Aguado *et al.* (2002a) y en EsperOnto (2003) pero, de entre todos los criterios incluidos, es el de c*apturabilidad* del CES, que postula que "el esquema de anotación debe contener los distintos niveles de análisis del texto", el que introduce otro nuevo concepto, de especial relevancia para el estudio aquí presentado: los niveles estratificados de análisis lingüístico, que generan sus propios tipos de anotación, y que se presentan en la siguiente sección.

3.2 Niveles de anotación lingüística.

En Leech (1997a) se puede encontrar una lista de los diferentes niveles de anotación lingüística. Como afirma este autor, ningún corpus incluye todos ellos, sino sólo dos o, como mucho, tres. Algunos de estos niveles se encontraban sólo en un estado preliminar en aquel momento. En la siguiente subsección se presenta una lista más pequeña, pero más realista, de niveles de anotación (de lemas, morfosintáctica, sintáctica, semántica y de discurso). Estos niveles se incluyen en EAGLES (1996b).

3.2.1 Anotación de lemas.

La *anotación de lemas (lematización)* supone acompañar cada *token* léxico con su **lema***,* es decir, la palabra que uno buscaría realmente en un diccionario. En inglés, la anotación lemática se puede considerar redundante pero, en lenguas más flexivas, como el español, el índice de formas morfológicas por lema hace que la anotación de lemas sea una contribución muy valiosa para la extracción de información (Leech, 1997a).

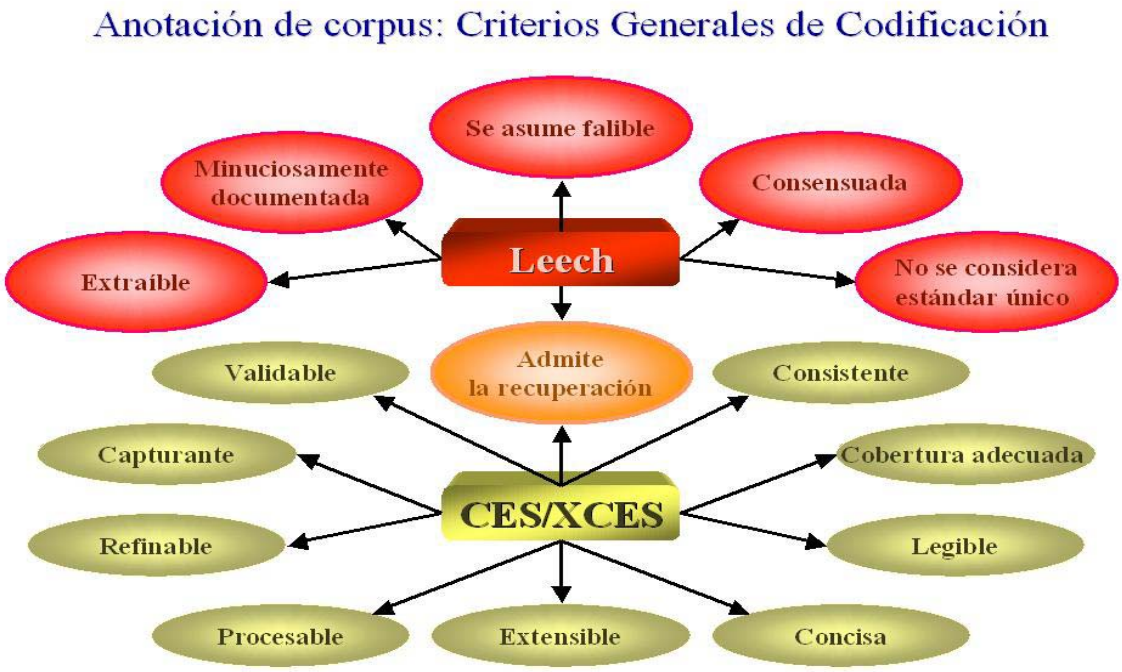

Ilustración 1: Criterios de Leech (1997a) y CES (1999) para la anotación de corpus.

3.2.2 Anotación morfosintáctica.

Éste es uno de los tipos de anotación más extendido en la Lingüística de Corpus, junto con la anotación sintáctica. La *anotación morfosintáctica*, *etiquetación POS* (del inglés, *part-of-speech*) o *etiquetación gramatical* es la anotación de la clase gramatical (por ejemplo, nombre, verbo, etc.) de cada *token* léxico en un texto[1], junto con (opcionalmente) la anotación de su análisis morfológico. Como se afirma en McEnery y Wilson (2001), la información POS constituye una base esencial para otras formas de análisis, como el sintáctico y el semántico. En la actualidad, una computadora puede llevar a cabo esta tarea con un alto grado de precisión sin intervención manual; sin embargo, todavía existen algunos problemas. La desambiguación de homógrafos, la identificación de expresiones idiomáticas, secuencias y compuestos o la separación de formas contraídas son algunas de las irregularidades que debe abordar un anotador en este nivel, ya que no se puede establecer una correspondencia biyectiva entre las unidades (palabras) ortográficas y las unidades morfosintácticas (Leech, 1997b). Soluciones para estos problemas se pueden encontrar en McEnery y Wilson (2001) y Leech (1997b) y, más concretamente para el español, en Pino y Santalla (1996).

3.2.3 Anotación sintáctica.

Una vez que se han identificado las categorías morfosintácticas de un texto, la *anotación sintáctica* añade la anotación de las relaciones sintácticas, en un nivel superior, entre estas categorías (determinadas, por ejemplo, mediante estructuras de frase o análisis de dependencias). Los anotadores de este nivel emplean distintos esquemas de análisis sintáctico. Según McEnery y Wilson (2001), estos esquemas difieren en:

§ los diferentes tipos de componentes de partida (generalmente, el número de etiquetas del etiquetario POS).

§ El modo en que se pueden combinar los componentes entre sí.

§ La gramática seguida para analizar y anotar el texto.

[1] En otras palabras, un sistema de etiquetación POS da respuesta a las preguntas a) ¿Cómo dividir el texto en *tokens* léxicos (palabras) con entidad propia?, b) ¿Cómo elegir el conjunto de etiquetas (= el conjunto de categorías que se aplicará a los *tokens* léxicos)? y c) ¿Cómo elegir qué etiqueta se debe aplicar a cada palabra (*token*)?.

3.2.4 Anotación semántica.

Como se afirma también en McEnery y Wilson (2001), es posible distinguir dos tipos principales de anotación semántica, relacionados con:

1. Las relaciones semánticas entre elementos del texto (es decir, los agentes, pacientes y participantes de acciones concretas). Este tipo de anotación apenas ha comenzado a aplicarse.
2. Las características semánticas de las palabras del texto, esencialmente la anotación de los significados de las palabras de una forma u otra. No hay un acuerdo universal en el ámbito de la semántica sobre qué características de las palabras se deben anotar[2].

Aunque ya se han planteado algunas recomendaciones preliminares sobre codificación léxico-semántica, EAGLES (1999) no ha publicado todavía recomendación alguna para la anotación semántica de corpus (como tal); sin embargo, para el segundo tipo de anotación semántica aludido, Schmidt (1988) ha propuesto un conjunto de criterios de referencia –mencionados en Wilson y Thomas (1997)– para diseñar un sistema de anotación de corpus basado en campos semánticos[3]. Estos criterios son:

1. *Debe tener sentido en términos lingüísticos o psico-lingüísticos*. Se sabe por experimentos psico-lingüísticos que en la mente existen unas categorías cognitivas o mentales básicas (Ungerer y Schmid, 1996). Hasta hoy, se ha logrado cierto consenso generalizado acerca de algunas de estas categorías básicas, que conocemos a través de las ciencias cognitivas (neuropsicología, lingüística cognitiva y otras) como son los colores, las partes del cuerpo, la topografía, etc.; sin embargo, queda aún un gran número de categorías por determinar hasta abarcarlas exhaustivamente. En cualquier caso, se debe evitar una abstracción excesiva.
2. *Debe ser posible dar explicaciones exhaustivas del vocabulario del corpus, no sólo de una parte de él*. Si no se puede clasificar un término en el sistema de anotación existente, entonces es necesario refinarlo.
3. *Debe ser suficientemente flexible como para tener en cuenta aquellos añadidos o retoques que sean necesarios para adaptarse a diferentes periodos, lenguas, registros o repositorios textuales*. Hoy día, debido a la cada vez mayor especialización del conocimiento, el tratamiento de textos especializados puede requerir una subclasificación considerablemente más detallada del dominio en cuestión que en un área más genérica. Es el caso de campos como los relacionados con la informática, el comercio, etc.
4. *Debe funcionar en un nivel de granularidad adecuado* –relacionado con el criterio (3). La determinación del nivel de granularidad adecuado para un sistema de anotación es una cuestión siempre abierta y depende parcialmente de los objetivos del usuario final. Por eso se plantea el siguiente criterio.
5. *Debe poseer, cuando sea apropiado, una estructura jerárquica*. Si una categoría semántica tiene una estructura jerárquica, basada en niveles cada vez más generales de relación entre términos, el usuario puede mirar todos los niveles y decidir cuáles debe emplear, simplemente subiendo o bajando de nivel en la jerarquía.
6. *Debe seguir un estándar, si es que existe*. Un sistema riguroso de categorías, incluso aunque sea el resultado de un trabajo consensuado, puede ser rechazado por muchos investigadores. Sin embargo, un estándar en este nivel podría identificar, como lo han hecho las recomendaciones EAGLES en otros niveles, un amplio marco de principios y categorías principales. que facilitara la comparación de resultados y, a la vez, podría ser modificado según las distintas necesidades individuales[4].

3.2.5 Anotación del discurso.

Éste es el tipo de anotación menos frecuente (en los corpus existentes hasta ahora). Se pueden encontrar dos tipos de enfoques en este nivel. El *enfoque de Stenström* (McEnery y Wilson, 2001) se basa en lo que ella llama *etiquetas de discurso*, deducidas empíricamente a partir del análisis inicial de una submuestra de un corpus. Incluyen categorías como "disculpas" (por ejemplo, *lo siento*, *perdone*) o "saludos" (por ejemplo, *hola*, *buenas noches*) y se usan para marcar elementos cuya función en el discurso tiene que ver principalmente con la gestión del mismo más que con su contenido proposicional.

---

[2] Véanse, por ejemplo, las controversias en las reuniones de las iniciativas SENSEVAL (Kilgarriff, 1998; Kilgarriff y Rosenzweig, 2000).

[3] Un **campo semántico** (a veces llamado campo conceptual, dominio semántico o dominio léxico) es una construcción teórica que agrupa palabras que están relacionadas por estar conectadas –en algún nivel de generalidad– con el mismo concepto mental (Wilson y Thomas, 1997).

[4] De nuevo se deben mencionar las iniciativas SENSEVAL, que revelan la demanda de estandarización semántica en el campo de la desambiguación de significados de palabras (Kilgarriff, 1998; Kilgarriff y Rosenzweig, 2000).

Este primer enfoque no ha llegado a extenderse en la Lingüística de Corpus. Por el contrario, el segundo enfoque, el de la referencia pronominal o *anotación anafórica* considera la *cohesión*[5] como un factor crucial de los procesos de lectura, producción y comprensión del discurso. Un exponente claro de este enfoque es el esquema UCREL de anotación del discurso. Otros esquemas de anotación anafórica son el de De Rocha, el de Gaizauskas y Humphries y el de Botley (Garside *et al.*, 1997).

Gracias a la iniciativa EAGLES de la UE, para algunos de estos niveles de anotación mencionados, se ha logrado un consenso acerca del grado y del modo en el que se debe anotar. Veamos a continuación brevemente en qué consiste esta iniciativa.

3.3 Recomendaciones EAGLES para la anotación (de corpus).

La iniciativa EAGLES proporciona una serie de recomendaciones, recogidas en un conjunto de documentos sobre buenas prácticas para anotación. Estas directivas comparten algunos principios (EAGLES, 1996b; 1996c):

1. Hacen uso de un formalismo atributo-valor.
2. No se ciñen a una jerarquía atributo-valor estricta (en términos de herencia monótona).
3. Usan tres subniveles de restricción (obligatorio, recomendado y opcional) para definir qué es aceptable según estas pautas:
   - § Las **anotaciones obligatorias** son de necesaria inclusión si se desea que el esquema de anotación para ese nivel cumpla las recomendaciones EAGLES.
   - § Las **anotaciones recomendadas** no son obligatorias, aunque no es conveniente su exclusión. Si ciertos atributos o valores recomendados se dan en una lengua concreta, estas directrices aconsejan que el conjunto de etiquetas de esa lengua las codifique.
   - § Las **anotaciones opcionales** son específicas de una (serie de) lengua(s) o una aplicación de ingeniería del lenguaje y se incluirán sólo en esos casos.

Veamos qué atributos y valores se consideran obligatorios y opcionales en EAGLES para las recomendaciones de anotación morfosintáctica (EAGLES, 1996b) y sintáctica (EAGLES, 1996c), las únicas hechas públicas hasta ahora.

[5] La **cohesión** (Halliday y Hasan, 1976) es el vehículo por el que se interconectan de una forma precisa los elementos de un texto o discurso, mediante el uso anafórico de pronombres (Palomar *et al.* 2001), la repetición, etc.

3.3.1 Nivel morfosintáctico.

- § Sólo se considera obligatorio un atributo: aquél de las principales categorías gramaticales (POS): N –nombre–, V –verbo–, AJ–adjetivo–, etc..
- § Atributos como tipo –común/propio–, género, número o caso se recomiendan para los nombres; además de los de persona, género, número, tiempo, voz, etc. para los verbos; y grado, género, número y caso para los adjetivos.
- § Los atributos y valores opcionales, o las extensiones especiales como se llaman en este documento, se subdividen en:
  - » Atributos y valores genéricos: por ejemplo, contabilidad –contable/masa–, para los nombres; aspecto –perfectivo/imperfectivo– separabilidad –no separable/separable– etc. para los verbos.
  - » Atributos y valores específicos de una lengua: por ejemplo, como sucede en danés con la distinción –definido/indefinido/ no marcado– para los nombres .

3.3.2 Nivel sintáctico.

- § Para este nivel, no se propone como obligatorio ningún atributo o valor, puesto que las anotaciones sintácticas pueden adoptar distintas formas, según la gramática en la que se basen (por ejemplo, gramática de estructura de frase, gramática de dependencias o gramática funcional).
- § Si se adopta una anotación basada en estructura de frase (no se detalla nada para otros casos) se recomienda el uso de las siguientes categorías: oración, cláusula, sintagma nominal, sintagma adjetivo, sintagma adverbial y sintagma preposicional.
- § Los ejemplos de anotaciones opcionales incluyen el marcado de tipos de oraciones (interrogativa, exhortativa, etc.), la anotación funcional de sujetos y objetos y la identificación de subtipos semánticos de componentes como los sintagmas adverbiales.

4 Integración de paradigmas: un ejemplo.

Como ya se ha mencionado, el objetivo de este trabajo es presentar la complementariedad de las anotaciones lingüística y ontológica para la Web Semántica. El fin del proyecto que presentamos, *ContentWeb*, es la creación de una plataforma basada en ontologías e integrada en *WebODE* (2003) que permita a los usuarios hacer consultas a aplicaciones de comercio electrónico usando lenguaje natural y también recuperar información de

```
<contentWeb:FilmReview>
<contentWeb:text>Tras cinco años de espera y después de
        muchas habladurías, llega a nuestras pantallas la película
        más esperada de los últimos tiempos.</contentWeb:text>
</contentWeb:FilmReview>

<!-- Morpho-syntactic annotation excerpt -->

<morphAnnot:Word rdf:ID="1_16">
<morphAnnot:surface_form>la</morphAnnot:surface_form>
<morphAnnot:TradAnnot rdf:about="#trad_ann_info_1_16"/>
<morphAnnot:MBTAnnot rdf:about="#mbt_ann_info_1_16"/>
<morphAnnot:ConstrAnnot rdf:about="#constr_ann_info_1_16"/>
</morphAnnot:Word>

<morphAnnot:TradAnnot rdf:ID="trad_ann_info_1_16">
<trad:tag> ARTDFS </trad:tag>
<morphAnnot:lemma> el </morphAnnot:lemma>
</morphAnnot:TradAnnot>

<morphAnnot:MBTAnnot rdf:ID="mbt_ann_info_1_16">
<mbt:tag> TDFS0 </mbt:tag>
<morphAnnot:lemma> el </morphAnnot:lemma>
</morphAnnot:MBTAnnot>

<morphAnnot:ConstrAnnot rdf:ID="constr_ann_info_1_16">
<constr:tag> DET </constr:tag>
        <constr:genus>FEM</constr:genus>
        <constr:numerus>SG</constr:numerus>
<morphAnnot:lemma>la</morphAnnot:lemma>
<constr:synfunction>DN&gt;</constr:synfunction>
</morphAnnot:ConstrAnnot>
```

Ilustración 2: Anotación morfosintáctica del artículo "la".

manera automática a partir de documentos web anotados con información ontológica y lingüística. Los objetivos de *ContentWeb* se pueden enunciar como sigue:

1. Construcción semiautomática de ontologías en los dominios del comercio electrónico y del ocio, reutilizando ontologías, iniciativas y estándares internacionales de comercio electrónico ya existentes.
2. Elaboración de *OntoTag*, un modelo y entorno para la anotación híbrida – lingüística y ontológica– de documentos web.
3. Desarrollo de *OntoConsult*, una interfaz en lenguaje natural basada en ontologías.
4. Creación de *OntoAdvice*, un sistema basado en ontologías para consultar y recuperar información a partir de documentos web anotados en el dominio del ocio y los espectáculos.

Para conseguir el segundo objetivo (elaboración de un modelo y entorno para la anotación híbrida de documentos web – *OntoTag*) una de las pruebas realizadas es la anotación, en los lenguajes XML y RDF(S), de una muestra de nuestro corpus de ocio y espectáculos, que ha sido extraído de diversos portales de Internet dedicados a estos temas. A continuación, presentamos la frase *"Tras cinco años de espera y después de muchas habladurías, llega a nuestras pantallas la película más esperada de los últimos tiempos."* como ejemplo del resultado obtenido tras la anotación en RDF(S) para los tres primeros niveles (Ilustraciones 2, 3 y 4).

En el nivel morfosintáctico (Ilustración 2) se da a cada palabra o elemento léxico un *Identificador Uniforme de Recurso* (Uniform Resource Identifier o URI). Se presenta la anotación morfosintáctica del artículo "la" según tres conjuntos de etiquetas diferentes. A cada conjunto de etiquetas se le ha asignado una clase distinta en el espacio de nombres, *morphAnnot*: *TradAnnot* (CRATER), *MBTAnnot* (MBT, 2002) y *ConstrAnnot* (Tapanainen y Järvinen, 1997).

En el nivel sintáctico (Ilustración 3), a cada relación entre elementos morfosintácticos se le da un nuevo URI para que se pueda referenciar en relaciones de nivel más alto o por otros niveles del modelo de anotación (es decir, *<synAnnot:Chunk rdf:ID="1_510">*). En la figura se ha incluido la anotación de la frase *"la película más esperada de los últimos tiempos"*.

```
<synAnnot:Chunk rdf:ID="1_510">
<synAnnot:synfunction>NP</synAnnot:synfunction>
<synAnnot:hasChild rdf:about="#1_21">los</synAnnot:hasChild>
<synAnnot:hasChild rdf:about="#1_22">últimos</synAnnot:hasChild>
<synAnnot:hasChild rdf:about="#1_23">tiempos</synAnnot:hasChild>
</synAnnot:Chunk>

<synAnnot:Chunk rdf:ID="1_511">
<synAnnot:synfunction>PP</synAnnot:synfunction>
<synAnnot:hasChild rdf:about="#1_20">de</synAnnot:hasChild>
<synAnnot:hasChild rdf:about="#1_510"> los últimos tiempos
        </synAnnot:hasChild>
</synAnnot:Chunk>

<synAnnot:Chunk rdf:ID="1_512">
<synAnnot:synfunction>AdjP</synAnnot:synfunction>
<synAnnot:hasChild rdf:about="#1_18">más</synAnnot:hasChild>
<synAnnot:hasChild rdf:about="#1_19">esperada</synAnnot:hasChild>
<synAnnot:hasChild rdf:about="#1_511">de los últimos tiempos
        </synAnnot:hasChild>
</synAnnot:Chunk>

<synAnnot:Chunk rdf:ID="1_513">
<synAnnot:synfunction>NP</synAnnot:synfunction>
<synAnnot:hasChild rdf:about="#1_16">la</synAnnot:hasChild>
<synAnnot:hasChild rdf:about="#1_17">película</synAnnot:hasChild>
<synAnnot:hasChild rdf:about="#1_512">más esperada de los últimos
        tiempos </synAnnot:hasChild>
</synAnnot:Chunk>
```

Ilustración 3: Anotación sintáctica de "la película más esperada de los últimos tiempos" en RDF(S).

```
<!-- Semantic annotation excerpt -->

<onto:PremiereEvent rdf:ID="_anon27">
    <semSynAnnot:includes rdf:about="#1_13">llega</semSynAnnot:includes>
    <semSynAnnot:includes rdf:about="#1_509">a nuestras pantallas</semSynAnnot:includes>
    <onto:hasFilm rdf:about="#_anon30"/>
</onto:PremiereEvent>

<onto:Film rdf:ID="_anon30">
      <semAnnot:includes rdf:about="#1_18">película</semAnnot:includes>
      <onto:comment rdf:about="#_anon40">
      <onto:comment rdf:about="#_anon41">
</onto:Film>

<onto:ControversialFilm rdf:ID="_anon40">
     <semSynAnnot:includes rdf:about="#1_506">después de muchas habladurías</semSynAnnot:includes>
</onto:ControversialFilm>

<onto:AwaitedFilm rdf:ID="_anon41">
     <semSynAnnot:includes rdf:about="#1_503">Tras cinco años de espera</semSynAnnot:includes>
     <semSynAnnot:includes rdf:about="#1_512">más esperada de los últimos tiempos</semSynAnnot:includes>
</onto:ControversialFilm>

<onto:Film rdf:about="#_anon30">
     <semSynAnnot:includes rdf:about="#3_507">El Señor de los Anillos</semSynAnnot:includes>
     <onto:filmTitle>El Señor de los Anillos</onto:filmTitle>
</onto:Film>
```

Ilustración 4: Anotación semántica de "Tras cinco años de espera y después de muchas habladurías, llega a nuestras pantallas la película más esperada de los últimos tiempos."

En el nivel semántico (Ilustración 4) se anotan algunos componentes, ya anotados en niveles inferiores, con referencias semánticas a los conceptos, atributos y relaciones determinados por nuestra ontología (de dominio), implementada en el lenguaje DAML+OIL. El equipo lingüístico de nuestro proyecto está buscando más elementos susceptibles de anotación semántica. La parte discursiva y pragmática de *OntoTag* no se ha abordado todavía, por eso no se incluyen estos niveles en el ejemplo, aunque se espera seguir las teorías apuntadas en Mann y Thomson (1988) y Álvarez de Mon (2003).

5 Las ventajas del modelo integrado.

Como se muestra en el ejemplo anterior, parece que la IA y la Lingüística de Corpus, lejos de ser irreconciliables, pueden unirse para dar lugar a un modelo de anotación integrada. Este esquema de anotación conjunta sería muy útil y valioso en el desarrollo de la Web Semántica. Además, se beneficiaría de los resultados de ambas disciplinas de muchas formas: primero, en el nivel semántico; segundo, en el resto de niveles. Por último, se dedican subsecciones concretas a la reutilización y la multifuncionalidad.

5.1 En el nivel semántico.

Veamos los beneficios que para el nivel semántico conlleva la utilización de un modelo de anotación híbrida, desde el punto de vista lingüístico y desde el ontológico.

5.1.1 La anotación basada en ontologías desde el punto de vista lingüístico.

El primer resultado de nuestro trabajo es que el uso de ontologías como base para un esquema de anotación semántica concuerda y cumple perfectamente los criterios planteados por Schmidt (1988) y mencionados en el apartado 3.2.4. Claramente, su estructura jerárquica (por regla general) cumple por sí misma el criterio (5), y, como efecto colateral, los criterios (2) y (4), dado que una ontología puede crecer horizontalmente (en extensión) y verticalmente (en profundidad). El criterio (3) también se satisface, dado que siempre podemos profundizar en el nivel de detalle de los conceptos de la ontología según periodos, lenguas, registros o repositorios textuales específicos. Las ontologías son, por definición, consensuadas y, por tanto, están más cerca de convertirse en estándar que muchos otros modelos de conocimiento, como requiere el criterio (6). En lo referente al criterio (1), muchos de los grupos que desarrollan ontologías se caracterizan por un enfoque interdisciplinar que combina la Informática, la Lingüística y, a veces, la Filosofía; por tanto, un enfoque basado en

ontologías debería tener sentido además en términos lingüísticos.

5.1.2 Las anotaciones lingüísticas desde un punto de vista ontológico.

El principal inconveniente para que los investigadores de IA adopten un modelo de anotación motivado lingüísticamente radica en el hecho de que (subsección 3.2.4) "no hay un acuerdo universal en la semántica sobre qué características de las palabras se deben anotar" ni sobre el criterio (1) de Schmidt (1988): "todavía falta por determinar un conjunto exhaustivo de categorías". Pero los investigadores en ontologías están intentando llenar este hueco con iniciativas como UNSPSC (2002) o RosettaNet (2002) para dominios específicos (por ejemplo, comercio electrónico). De cualquier modo, las anotaciones lingüísticas en el nivel semántico son más ambiciosas y potencialmente más amplias que las basadas exclusivamente en ontologías y, por ello, probablemente más complejas y con un coste mucho mayor en horas/persona. Establecer un enlace entre la anotación semántica y la del discurso y la construcción de textos siguiendo el enfoque RST (Mann y Thomson, 1988), que ya se ha aplicado en la generación de textos, parece una mejora lingüística bastante prometedora.

Hasta ahora, hemos visto cómo las ontologías pueden encajar en la anotación semántica de textos; veamos en las siguientes subsecciones cómo las anotaciones lingüísticas de todos sus niveles pueden mejorar el tratamiento automático de documentos de la Web Semántica.

5.1.3 El significado no está sólo en la semántica.

Como se afirma en Pulman (1995), todos los niveles lingüísticos interactúan estrechamente para determinar el significado de una oración, una declaración o una expresión completa. Por otra parte, aunque los componentes básicos de una expresión[6] sean los significados de las palabras, el significado de una expresión estará caracterizado no sólo por los significados de sus palabras, sino también por el orden de combinación que presenten. Dado que los modos de combinación están muy determinados por la estructura sintáctica de la lengua, necesitaremos reflejar el significado que aporta cada regla sintáctica subyacente en la expresión que se está analizando, es decir, la operación semántica que combina los significados de los hijos (analizados) para producir el significado del padre dentro del árbol sintáctico asociado a la expresión. Por consiguiente, *para determinar el significado de una expresión se necesita realizar su análisis sintáctico*. Dik (1989), Aguado y Pareja-Lora (2000) y Vargas-Vera *et al.* (2001) han confirmado la importancia de interrelacionar sintaxis y semántica. Por otro lado, Pulman (1995) también señaló la necesidad de una mayor integración entre la semántica oracional y las teorías de la estructura textual o discursiva, incluyendo aspectos como la composición/escenario del texto o diálogo, o sobre la intencionalidad de los hablantes. Así, *para ayudar a determinar el significado de una expresión en un texto, conviene llevar a cabo algún tipo de análisis pragmático explícito o implícito*. Por tanto, podemos llegar a la conclusión de que sería muy útil para la comunidad de la Web Semántica tener algún modelo de anotación que permita no sólo anotar y hacer explícito el nivel semántico, sino también los otros niveles, y que facilite el tratamiento automático de páginas web mediante su inclusión y anotación explícita en páginas de la Web Semántica.

5.1.4 Significado y anotación de lemas.

La lematización puede ayudar notablemente en el procesamiento semántico, por ejemplo, al facilitar la extracción de información para lenguas muy flexivas, como el español o el alemán (Kietz *et al.*, 2000), sobre todo cuando entran las ontologías en consideración: la anotación de lemas prepara el terreno para una anotación semántica (semi)automática basada en ontologías.

5.1.5 Significado y anotación morfosintáctica.

Muchos proyectos de extracción de información basados en ontologías usan algún tipo de análisis morfosintáctico (Martínez-Fernández y García-Serrano, 2002; Vargas-Vera *et al.*, 2001; Kietz *et al.*, 2000) como fase preliminar para el procesamiento semántico. Por ello, podemos considerar la etiquetación POS como un tipo de anotación preliminar, básica y necesaria, un primer paso hacia niveles de anotación superiores, como el sintáctico o el semántico. Como se afirma en la subsección 0, se deberían identificar y marcar algunos sintagmas nominales y otras secuencias de expresiones idiomáticas o paráfrasis (por ejemplo, "llega a nuestras pantallas", "El Señor de los Anillos"[7]) como unidades léxicas y anotarlos consecuentemente. Un modo hábil de lograr esto para el español se puede encontrar en Pino y Santalla (1996).

[6] Gran parte de la información contenida en una página web se presenta en forma suboracional (principalmente sintagmas nominales). De ahí que, en adelante, se prefiera el término "expresión".

[7] Ejemplos extraídos de nuestro corpus en el dominio del ocio.

5.1.6 Significado y anotación sintáctica.

De nuevo debemos mencionar a Vargas-Vera *et al.* (2001) y a Kietz *et al.* (2000), puesto que en sus proyectos hacen uso de algún tipo de análisis sintáctico cuando procesan documentos. Se consideran dos tipos de anotaciones sintácticas útiles desde un punto de vista semántico:

1. Las anotaciones opcionales de EAGLES como el tipo de oración, el sujeto y los complementos o los subtipos semánticos constituyentes (ver apartado 3.3.2).
2. Fenómenos sintácticos concretos de una lengua, como la identificación y marcado de verbos separables para el alemán.

5.1.7 Significado y anotación de discurso.

Dado que este nivel de anotación se abordará en etapas posteriores de nuestro proyecto, sólo podemos hacer notar la utilidad potencial de un esquema de anotación anafórica para extraer la cohesión del documento procesado.

5.2 Reutilización.

Como se ha señalado, en Vargas-Vera *et al.* (2001) y Kietz *et al.* (2000) se justifica la necesidad de un análisis sintáctico (superficial) para el procesamiento semántico de páginas web: la mayoría de los sistemas de extracción de información (además de otras aplicaciones de PLN) emplean alguna forma de análisis sintáctico superficial[8] para reconocer construcciones sintácticas o, en otras palabras, para identificar los grandes bloques sintácticos de una oración. Esto complica y reduce la velocidad del proceso de análisis semántico de un documento de la WWW. Aunque el proceso de creación y edición de una página podría parecer abrumador no debemos olvidar que existen algunas herramientas gratuitas para fines investigadores. De este modo, se reutilizan las herramientas junto con los resultados que producen y que se incluyen como anotaciones dentro de la página web (véase el ejemplo de la sección 4).

5.3 Multifuncionalidad.

Aunque gran parte de los beneficios mencionados se aplican a los sistemas de extracción de información, no son exclusivos de este tipo de aplicaciones de PLN. Dado que el modelo de anotación propuesto añade información lingüística explícita a cualquier tipo de documento, puede usarse para múltiples fines que requieran un análisis o procesamiento semántico (por ejemplo, traducción automática, recuperación de información, etc.).

[8] Es decir, sin generar un árbol de análisis completo para cada oración. Ese análisis sintáctico parcial tiene las ventajas de ser más rápido y robusto.

6 Conclusiones y trabajos futuros.

Hemos visto que, aunque los investigadores en IA están dedicando muchos esfuerzos a encontrar un modelo óptimo para la anotación semántica de páginas web, las décadas de trabajo y los resultados obtenidos en el campo de la Lingüística de Corpus sobre anotación de corpus han sido, de algún modo, ignorados por ellos, sobre todo en los niveles distintos del semántico. Hemos visto también que estos otros niveles lingüísticos aportan alguna información semántica que puede ayudar a que la computadora entienda páginas de la Web Semántica. En este artículo hemos presentado los distintos niveles lingüísticos en que se puede anotar un documento y se han mostrado los resultados de nuestra investigación, llevada a cabo para determinar cómo la anotación lingüística puede ayudar a que las computadoras comprendan el texto contenido en un documento –una página de la Web Semántica–, combinando los modelos de anotación semántica de la IA con las anotaciones propuestas para cada nivel lingüístico por la Lingüística de Corpus. Conviene tener en cuenta que, en esta primera fase de nuestro proyecto, el objetivo primordial no es la eficiencia, sino que, más bien, buscamos tratar los significados desde nuevos puntos de vista, que permitan una profundidad mayor de la conseguida hasta la fecha. La integración de estos dos enfoques (IA y LC) en los distintos niveles de anotación mencionados supone muchas ventajas para las aplicaciones de IA e Ingeniería Lingüística. Primero, los recursos lingüísticos serán más reutilizables: muchos proyectos que conllevan el uso de documentos (web) anotados semánticamente también analizan sintácticamente la información y, antes que eso, deben determinar la categoría gramatical asociada a cada palabra del documento. Introducir la anotación de estos dos niveles en el documento, reutilizando, por tanto, una de las herramientas ya desarrolladas con este fin, evita repetir innecesariamente todo este proceso de *tokenización* y análisis sintáctico (reutilizando la anotación). Dado que el análisis sintáctico, por ejemplo, es una tarea que lleva mucho tiempo, podemos tener una ventaja adicional, es decir, reducir nuestro tiempo global de procesamiento de páginas de la Web Semántica. La segunda ventaja principal es que el significado de una página con anotación semántica explícita se puede reforzar mediante la contribución de significado proporcionada por todos los niveles lingüísticos; como contrapartida, el análisis semántico también se puede beneficiar del extraordinario trabajo hecho hasta ahora sobre el

desarrollo de ontologías como modelos conceptuales consensuados.

Sin embargo, el principal inconveniente reside en las limitaciones impuestas por las tecnologías actuales: el proceso de obtener automáticamente páginas compactas, legibles y verificables es una tarea de muy difícil delimitación y especificación en toda su magnitud. Por otro lado, la inclusión de todos estos niveles de anotación en un documento web conlleva un evidente (aunque no exagerado) aumento del tiempo de descarga del documento desde la red, pero el trabajo hecho en nuestro laboratorio intenta aportar alguna luz sobre estos problemas.

Agradecimientos.

La investigación presentada en este artículo ha sido realizada con financiación del Ministerio Español de Ciencia y Teconología (MCyT) a través del proyecto ContentWeb: "PLATAFORMA TECNOLÓGICA PARA LA WEB SEMÁNTICA: ONTOLOGÍAS, ANÁLISIS DE LENGUAJE NATURAL Y COMERCIO ELECTRÓNICO" – **TIC2001-2745**. Asimismo, queremos dar las gracias al grupo de ontologías del Laboratorio de Inteligencia Artificial (LIA) de la UPM, por su ayuda con los aspectos ontológicos de nuestro trabajo.

Referencias.

AeroDAML (2002) http://ubot.lockheedmartin.com/ubot/hotdaml/aerodaml.html

Aguado, G. y Pareja-Lora, A. (2000) A competition model for the generation of complementation patterns in machine translation. *International Journal of Translation*, Vol. 12, Nº 1-2, Jan-Dec 2000. Bahri Publications. New Delhi, INDIA.

Aguado, G., Álvarez-de-Mon I., Gómez-Pérez, A., Pareja-Lora, A. y Plaza-Arteche, R. (2002a) A Semantic Web Page Linguistic Annotation Model. *Semantic Web Meets Language Resources. Technical Report WS-02-16*. American Association for Artificial Intelligence. AAAI Press. Menlo Park, California, E.E.U.U.

Aguado, G., Álvarez-de-Mon, I., Pareja-Lora, A. y Plaza-Arteche, R. (2002b) *OntoTag*: A Semantic Web Page Linguistic Annotation Model. *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation and Knowledge Markup*. Lyon, Francia.

Aguado, G., Álvarez-de-Mon, I., Pareja-Lora, A. y Plaza-Arteche, R. (2002c) RDF(S)/XML linguistic annotation of Semantic Web pages. *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002)*. COLING'2002. Taipei, Taiwan.

Álvarez-de-Mon I. (2003) *La cohesión del texto científico-técnico: un estudio contrastivo inglés-español.* Universidad Complutense de Madrid (en imprenta).

Benjamins, V.R., Fensel, D., Decker, S. y Gómez-Pérez, A. (1999) $(KA)^2$: Building Ontologies for the Internet: a Mid Term Report. *IJHCS, International Journal of Human Computer Studies*, 51: 687–712.

Berners-Lee, T. y Fischetti, M. (1999) *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. San Francisco: Harper.

Bray, T., Paoli, J. y Sperberg, C. (1998) *Extensible Markup Language (XML) 1.0.* W3C Recommendation. http://www.w3.org/TR/REC-xml

Brickley, D. y Guha, R.V. (2000) *Resource Description Framework (RDF) Schema Specification*. W3C Candidate Recommendation. http://www.w3.org/TR/PR-rdf-schema.

Buitelaar, P., Bryant, B., Ide, N., Lin, J., Pareja-Lora, A. y Wilcock, G. 2003. The Roles of Natural Language and XML in the Semantic Web. *Language and Linguistics* (en prensa).

CES (1999) http://www.cs.vassar.edu/CES/

COHSE (2002) http://cohse.semanticweb.org/

Dik, S.C. (1989) *The Theory of Functional Grammar*. Dordrecht: Foris Publications.

EAGLES (1996a) *EAGLES: Text Corpora Working Group Reading Guide.* EAGLES Document EAG-TCWG-FR-2. http://www.ilc.pi.cnr.it/EAGLES96/home.html

EAGLES (1996b) *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG--TCWG—MAC/R. http://www.ilc.pi.cnr.it/EAGLES96/home.html

EAGLES (1996c) *EAGLES: Recommendations for the Syntactic Annotation of Corpora*. EAGLES Document EAG--TCWG—SASG/1.8. http://www.ilc.pi.cnr.it/EAGLES96/home.html

EAGLES (1999) *EAGLES LE3-4244: Preliminary Recommendations on Semantic Encoding*, Final Report. http://www.ilc.pi.cnr.it/EAGLES/EAGLESLE.PDF

EsperOnto (2003) *Esperonto Services IST-2001-34373 Deliverable on Annotation*. http://www.esperonto.net/ (en edición).

Garside R., Fligelstone, S. y Botley, S. (1997) Discourse Annotation: Anaphoric Relations in Corpora. In Garside R., Leech, G. y McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London.

Gruber, R. (1993) A translation approach to portable ontology specification. *Knowledge Acquisition.* #5: 199-220.

Halliday, M. y Hasan, R. (1976) *Cohesion in English*. London: Longman.

Horrocks, I., Fensel, D., Harmelen, F., Decker, S., Erdmann, M y Klein, M. (2000) OIL in a Nutshell. *12th International Conference in Knowledge Engineering and Knowledge Management, Lecture Notes in Artificial Intelligence, 1–16*. Berlin, Germany: Springer-Verlag. http://www.cs.vu.nl/~ontoknow/oil/downl/oilnutshell.pdf

Horrocks, I. y Van Harmelen, F. (2001) *Reference description of the DAML+OIL ontology markup language.* Draft report, 2001. http://www.daml.org/2000/12/reference.html

Karp, R., Chaudhri, V. y Thomere, J. (1999) *XOL: An XML-Based Ontology Exchange Language*. Technical Report. http://www.ai.sri.com/~pkarp/xol/xol.html

Kent, R. (1998) *Conceptual Knowledge Markup Language (version 0.2).* http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Kent1/CKML.pdf

Kietz, J-U., Maedche, A. y Volz, R. (2000) A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. *Proceedings of the EKAW'00 Workshop on Ontologies and Text*. Juan-Les-Pines, France.

Kilgarriff, A. (1998) SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. *Proceedings of LREC*, 581–588. Granada, Spain. ftp://ftp.itri.bton.ac.uk/reports/ITRI-98-09.ps.gz

Kilgarriff, A. y Rosenzweig, J. (2000) English SENSEVAL: Report and Results. *Proceedings of LREC*. Athens, Greece. ftp://ftp.itri.bton.ac.uk/reports/ITRI-00-25.ps.gz

Lassila, O. y Swick, R. (1999) *Resource Description Framework (RDF) Model and Syntax Specification*. W3C Recommendation. http://www.w3.org/TR/PR-rdf-syntax

Leech, G. (1997a) Introducing corpus annotation. In Garside R., Leech, G. y McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Leech, G. (1997b) Grammatical tagging. In Garside R., Leech, G. y McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Luke S. y Heflin J. (2000) *SHOE 1.01. Proposed Specification*. SHOE Project. http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.htm

Mann, W y Thomson, S. (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, Vol.18, 3: 243–281.

Martínez-Fernández, P. y García-Serrano, A. M. (2002) Interacción persona-web empleando recursos lingüísticos. *Revista Iberoamericana de Inteligencia Artificial*, número 16 – Monografía: Interacción Persona-Ordenador: 55–65. http://tornado.dia.fi.upm.es/caepia/numeros/16/pmf.pdf

MBT (2002) http://ilk.kub.nl/~zavrel/tagtest.html

McEnery, A. M. y Wilson, A. (2001) *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

Motta, E., Buckingham Shum, S. y Domingue, J. (1999) Case Studies in Ontology-Driven Document Enrichment. *Proceedings of the 12th Banff Knowledge Acquisition Workshop*, Banff, Alberta, Canada.

Nirenburg, S. y Raskin, V. (2001) *Ontological Semantics (Draft)* http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/index-book.html.

OntoMat (2002) http://annotation.semanticweb.org/ontomat.html

OntoWeb (2002) http://ontoweb.aifb.uni-karlsruhe.de/About/Deliverables/D13_v1-0.zip

Palomar, M., Saiz-Noeda, M., Muñoz, R., Suárez, A., Martínez-Barco, P. y Montoyo, A. (2001) PHORA: A NLP System for Spanish. Alexander F. Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing*, Mexico-City: Springer-Verlag: 126-139.

Pino, M. y Santalla, P. (1996) http://www.cica.es/sepln96/sepln96.html

Pulman, S. G. (1995) http://cslu.cse.ogi.edu/HLTsurvey/ch3node7.html#SECTION35

RosettaNet (2002) *RosettaNet: Lingua Franca for eBusiness.* http://www.rosettanet.org/

Schmidt, K. M. (1988) Der Beitrag der begriffsorientierten Lexicographie zur systematischen Erfassung von Sprachwandel und das Begriffswörterbuch zur mhd. Epik. In Bachofer, W. (ed.) *Mittelhochdeutsches Wörterbuch in der Diskussion*. Tübingen: Max Niemeyer, 35–49.

SHOE (2002) http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html

Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Mädche, A., Schnurr, H.-P. y Studer, R. (2000) Semantic Community Web Portals. *WWW9 - Proceedings of the 9th International World Wide Web Conference*, 33(1-6): 473-491 (Special Issue). Amsterdam, Holanda: Elsevier.

Studer, R., Benjamins, R. y Fensel, D. (1998) Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering (DKE)*, Vol. 25 (1-2): 161-197.

Tapanainen, P. y Järvinen, T. (1997) A non-projective dependency parser. *Proceedings of the 5th conference on Applied Natural Language Processing*. Washington D.C.: Association for Computational Linguistics, 64–75.

Ungerer, F. y Schmid, H. J. (1996) *An Introduction to Cognitive Linguistics*. London: Longman.

UNSPSC (2002) *Universal Standard Products and Services Classification (UNSPSC).* http://www.unspsc.org/

Vargas-Vera, M., Motta, E., Domingue, J., Shum, S. B. y Lanzoni, M. (2001) Knowledge Extraction by Using an Ontology-based Annotation Tool. *Proceedings of the K-CAP'01 Workshop on Knowledge Markup and Semantic Annotation*, Victoria B.C., Canada.

WebODE (2003) http://delicias.dia.fi.upm.es/webODE/

Wilson, A. y Thomas, J. (1997) Semantic Annotation. R. Garside, G. Leech y A. M. McEnery, (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.