La ontología: una zona de interacción entre la Lingüística y la Documentación

Silvia Arano

Citación recomendada: Silvia Arano. *La ontología: una zona de interacción entre la Lingüística y la Documentación* [en linea]. "Hipertext.net", núm. 2, 2003. <http://www.hipertext.net> [Consulta: 12 feb. 2007]. .

1. Introducción
2. La lingüística y la documentación
   2.1. El concepto de reutilización (reusability)
   2.2. La lengua: objeto complejo para procesar
3. Aproximación a los recursos lingüísticos desde la lingüística
   3.1. Ontología
   3.2. Base de conocimiento
   3.3. Base de datos léxica
      3.3.1. 'Wordnet: a lexical database for english language'
   3.4. Lexicón computacional
   3.5. Tesauro
   3.6. Síntesis
4. Los recursos lingüísticos: aplicabilidad desde la documentación
5. A modo de conclusión
6. Bibliografía
7. Notas

1. Introducción

El objetivo [1] inicial del presente trabajo era estudiar el uso de las ontologías como recurso en la Lingüística. No obstante, al realizar el estudio de la producción bibliográfica sobre el tema se observó que la ontología era un recurso lingüístico que en muchas ocasiones se vinculaba o incluso se confundía con otros recursos, tales como las base de conocimiento, la base de datos léxica, el lexicón computacional y el tesauro. De ahí se mostró la necesidad de sistematizar y precisar terminológicamente estos diferentes recursos lingüísticos, así como el interés por determinar la utilidad de su aplicación al ámbito de la Documentación. Por consiguiente fue necesaria una reformulación de la propuesta inicial de investigación, que derivó en la proposición de los siguientes objetivos de este trabajo que presentamos aquí:

1. establecer diferencias y semejanzas entre los recursos lingüísticos
2. determinar si son aplicables en la Documentación.

En el presente trabajo el punto de partida es el uso impreciso de la denominación de ontología, base de conocimiento, base de datos léxica, lexicón computacional y tesauro en la bibliografía especializada donde se hace una referencia confusa a dichos recursos lingüísticos.

El artículo está estructurado en tres partes. En la primera, se establecen las relaciones entre la Lingüística y la Documentación basándonos en la lengua como elemento común en ambas prácticas, como también presentando el concepto de reutilización y la complejidad de la lengua como objeto para procesar informáticamente. En la segunda parte, se ofrece una descripción de los distintos recursos lingüísticos estudiados incluyendo, en el apartado de base de datos léxicas, el análisis de WordNet como recurso lingüístico cuya interpretación y utilización ha sido fuente de múltiples imprecisiones. En la tercera, se presenta la aplicabilidad de los recursos lingüísticos a la Documentación. Finalmente se presentan las conclusiones y perspectivas de investigación futuras.

2. La lingüística y la documentación

La lengua es el instrumento para la exploración y explotación de la información que utiliza el documentalista en los procesos de gestión de la información. Los documentos, continentes de la información, son estructuras lingüísticas donde la información se encuentra codificada a través de la lengua. El documentalista, en su función de intermediario entre la organización y la recuperación de la información, necesita acceder al contenido de los documentos para

interpretarlos utilizando un conocimiento léxico, sintáctico-semántico y contextual.

El Procesamiento del Lenguaje Natural (en adelante PLN), en estrecha relación con la Lingüística y las ciencias cognitivas, también incide en la práctica de la Documentación. El concepto de lenguaje natural varía de sentido si se observa desde la Lingüística o desde la Documentación. Para la primera, el *lenguaje natura* l se entiende en contraposición al *lenguaje formal* creado artificialmente para describir a los lenguajes naturales. En cambio para la segunda, el *lenguaje natural* se contrapone *al lenguaje controlado* en referencia a las fuentes de donde se extraen los términos para indizar los documentos. La indización puede tomar los descriptores de partes propias del documento (título, resumen o texto) *lenguaje natural* o de tesauros, listas de encabezamiento de materia, etc. *lenguaje controlado* . Sin embargo, tanto para la Lingüística como para al Documentación, la materia de trabajo es la lengua, entendida como un *'sistema de signos que utiliza una comunidad para comunicarse oralmente o por escrito* '. (*Lema : diccionario de la lengua española* 2001: 1057)

El PLN se ocupa de la comprensión y la interpretación del lenguaje natural en el marco de aplicaciones concretas. Gran parte de las aplicaciones del PLN son implementables en ambos ámbitos disciplinarios, ejemplos de ello son la traducción automática, la indización automática de textos, la interacción hombre-ordenador en lenguaje natural, la elaboración automática de resúmenes y la extracción y recuperación de información.

2.1. El concepto de reutilización (reusability)

El concepto de *reutilización,* que proviene de la Informática, se adopta en la Lingüística a partir de la década de los 80's, cuando distintos grupos de investigación (europeos, americanos y japoneses) plantearon proyectos y actividades con el fin de investigar el léxico de una forma más sistemática. Para emprender este tipo de actividades comunes se debe tener en cuenta la reutilización de los recursos lingüísticos con el fin de ahorrar costes.

La *reutilización* tiene dos perspectivas de interpretación. Por un lado, refiere a reutilizar recursos lingüísticos existentes, aunque su soporte no sea informático, los que son usados por aplicaciones de PLN con el fin de elaborar nuevos recursos. Por el otro, remite a la construcción de recursos lingüísticos, ya sea creándolos o utilizando fuentes que se hayan generado a partir de otros recursos ya existentes, que a su vez, pueden reutilizarse en distintas aplicaciones, marcos teóricos y tipos de usuarios (humanos y/o mecánicos).

La posibilidad de *reutilización* de los recursos lingüísticos no sólo es importante para la propia Lingüística sino que también lo es para otras áreas del conocimiento, incluyendo la Documentación. Dada la proliferación de tecnologías aplicadas al lenguaje, la Documentación debe actualizar su práctica, automatizar los procesos de gestión de información y agilizar las tareas documentales en estrecha colaboración con los recursos generados en la Lingüística, para consolidar tanto una práctica interdisciplinar como transdisciplinar.

2.2. La lengua: objeto complejo para procesar

La lengua, como objeto de procesamiento informático, muestra su complejidad cuando los programas informáticos deben resolver situaciones de ambigüedad en cualquiera de los niveles lingüísticos (léxico, semántico o sintáctico). La ambigüedad en la lengua para los sistemas informáticos supone un desafío, pues deben hacer explícito todo el conocimiento que es necesario para la interpretación adecuada de las producciones lingüísticas.

Alonso Martín (2001) permite acercarnos a la problemática de la ambigüedad de la lengua a través de los siguientes ejemplos donde sólo parece cambiar el sujeto:

- ' *Els pingüins poden nedar però no volen* ' (Los pingüinos pueden nadar pero no volar) [2] *'Els nens poden nedar però no volen'* (Los niños pueden nadar pero no quieren) Sin embargo para ambos casos, el verbo *volen* en catalán, se corresponde con la forma conjugada de la tercera persona del plural del modo indicativo pero de distintos verbos, *volar* (volar) y *voler* (querer) respectivamente. Debido a ello, la interpretación habitual sería que: los pingüinos *pueden* nadar, pero no pueden volar; y los niños pueden nadar, pero no *quieren* hacerlo.

Las siguientes suposiciones dan las pautas para explicar la distinta interpretación de cada frase:

- los pingüinos son aves, tienen alas, pero no lo suficientemente desarrolladas para volar
- los pingüinos son aves, pero tienen la cualidad de nadar
- los niños no pueden volar por sí mismos (sí, si lo hacen en un avión, por ejemplo)
- decir que los niños pueden nadar pero no vuelan, solamente es posible en un contexto muy restringido
- decir que un pingüino puede nadar pero no lo quiere hacer, también es posible, pero nuevamente es dependiente de estar en un discurso muy específico.

Proporcionar información relevante para un programa informático que procese esta serie de suposiciones que distingan

volen , en el sentido de volar, y *volen,* en el sentido de querer, implica transferir a este programa informático el conocimiento y el funcionamiento de la lengua, así como especificar formalmente este conocimiento para resolver cualquier tipo de ambigüedad.

3. Aproximación a los recursos lingüísticos desde la lingüística

El primer paso para el procesamiento informático del conocimiento lingüístico es la representación formal de dicho conocimiento, como indicábamos en punto 2.2. De los múltiples recursos creados para representar la información lingüística se proponen analizar: la ontología, la base de conocimiento, la base de datos léxica, el lexicón computacional y el tesauro, con el fin precisar sus denominaciones.

3.1. Ontología

'Large ontologies such as WordNet ...' (Fensel et al. 2001: 39)

Una ontología es una representación formal del conocimiento donde los conceptos, las relaciones y las restricciones conceptuales son explicitadas mediante formalismos en un determinado dominio. Su función más frecuente en la Lingüística es de apoyo para sistemas de Traducción Automática Basada en el Conocimiento y para la Terminografía (práctica de la Terminología). En ambos casos, la ontología es una representación formal y explícita de la estructura conceptual del campo sobre el que se trabaja. Este recurso lingüístico incluye como mecanismo de inferencia a la herencia, que implica una economía en la codificación de la información: los conceptos superiores transmiten sus características a los conceptos inferiores.

Por consiguiente, la ontología es uno de los módulos asociados a un sistema de conocimiento donde su función es la de apoyo semántico para las unidades léxicas, es decir que las unidades léxicas son descritas como objetos lingüísticos en una base de datos léxica y son relacionadas con una jerarquía conceptual localizada en una ontología.

3.2. Base de conocimiento

Las bases de conocimientos léxicos, también llamadas ontologías o tesaurus computacionales ... ' (Gamallo Otero 2000: 1)

Una base de conocimiento es una forma avanzada de base de datos que no solo pretende almacenar, recuperar y modificar grandes cantidades de información, sino también, plasmar elementos de conocimiento (generalmente en forma de hechos y reglas de inferencia), así como la forma en que éste ha de ser utilizado.

Este recurso lingüístico es un modelo de un mundo/empresa/sección de la realidad, como declara Mylopoulos (1982) citado por Pérez Hernández (2002), en donde se considera el mundo/universo como una colección de individuos o entidades y el conjunto de relaciones que existen entre esos individuos. La colección de individuos, más las relaciones entre éstos, constituye un estado, cuyos cambios causan la creación o modificación de individuos o de las relaciones entre ellos.

La base de conocimiento utiliza al *esquema de representación* como notación precisa para representar el conocimiento que contiene. El tipo de *esquema de representación* de mayor difusión son las redes semánticas que estructuran sus datos en nodos que simbolizan a los conceptos unidos por arcos que representan las relaciones conceptuales. La red semántica también utiliza la herencia como método de inferencia, donde los nodos inferiores heredan las características de los nodos superiores permitiendo así una economía de codificación.

Los tres tipos tradicionales de redes semánticas son: las redes IS-A, los grafos conceptuales y las redes de marcos (frames). El esquema basado en marcos (frames) es el más explorado y utilizado por los investigadores dada su mayor flexibilidad, y en mayor o menor medida, generan las ontologías de conceptos.

En Lingüística, siguiendo a Feliu et al. (2002), Moreno Ortiz (2001), Cabré et al. (2004), la base de conocimiento se concibe como un recurso modular que integra recursos tales como bases textuales, bases bibliográficas, bases terminológicas y ontologías.

3.3. Base de datos léxica

'WordNet is an online lexical database ...' (Miller 1995: 39)

Una base de datos léxica es un sistema de almacenamiento de información lingüística organizada según un determinado *modelo de datos* que posibilita el almacenamiento, recuperación y modificación de los mismos. El *modelo de datos*

puede tener una estructura jerárquica, de red o relacional, esta última la de más amplia difusión. Una base de datos léxica tiene la función de responder a consultas sobre los datos que contiene, ya sea desde prestaciones propias o a partir de aplicaciones externas, permitiendo la reutilización de la información contenida. El comportamiento de una base de datos léxica es pasivo, pues las operaciones sobre sus datos son realizadas por aplicaciones que deben ser iniciadas explícitamente.

Formalmente las bases de datos no están planeadas para guardar información compleja sino grandes volúmenes de datos, con lo cual su aplicación en la representación de la información léxica dificulta la visión de conjunto debido a la tendencia a la *atomicidad* de los datos almacenados. El concepto de *atomicidad* implica que un elemento sea atómico (o escalar), cuando no puede fragmentarse en partes más pequeñas. Por ejemplo, en la codificación de un número telefónico, si se opta por codificar la información en tres valores separados --prefijo internacional, prefijo interprovincial y número del abonado--, se posibilita una gestión impensable en caso de que todo el número se tomara como un único valor. Si el número telefónico es segmentado según los distintos prefijos que lo componen, un programa de comunicaciones que marque el número y establezca la conexión en forma automática puede utilizarlo. Por tanto, el concepto de *atomicidad* no es consecuente con las características de la información léxica.

Las bases de datos léxicas son utilizadas en la Lingüística como fuentes de información léxica a reutilizar por otros recursos, por ejemplo un lexicón computacional o una base de datos terminológica.

3.3.1. 'Wordnet: a lexical database for english language'

WordNet [3] es una base de datos léxica diseñada sobre la base de las teorías psicolingüísticas acerca del lexicón mental, cuya finalidad era agilizar las búsquedas de los diccionarios en línea para la lengua inglesa.

Esta base de datos léxica se construye sobre la base de las categorías sintácticas de nombre, verbo, adjetivo y adverbio. Dichas categorías se organizan en distintas estructuras léxicas: los nombres en jerarquías léxicas sobre la base de relaciones de hiponimia y meronimia; los verbos en base a relaciones de implicación (*entailment*), y finalmente, los adjetivos y adverbios se organizan como hiperespacios N-dimensionales. Sin embargo, este tipo de organización produce una redundancia de información en los casos en que una unidad léxica pertenece a más de una categoría.

WordNet se basa en el supuesto teórico de *matriz léxica* (*lexical matrix*) integrada por los elementos de *forma léxica (word form)* y *significado léxico (word meaning)* que se corresponden con la expresión física que se escribe o pronuncia, por un lado, y con el concepto lexicalizado que se expresa por medio de una forma léxica, por otro. La*forma léxica* (*word form*) se corresponde con la designación *de* unidad léxica que se está utilizando en el desarrollo del presente artículo.

Table 1

Illustrating the Concept of a Lexical Matrix:

F_1 and F_2 are synonyms; F_2 is polysemous

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
⋮				⋱	
M_m					$E_{m,n}$

(Fellbaum et al. 1993: 4)

En la *matriz léxica* el encabezamiento de las columnas (F1) corresponde a las unidades léxicas de una lengua y el encabezamiento de las filas (M1) a los conceptos. Una entrada en una celda de la matriz (E1,1) implica que esa forma (F1) puede ser utilizada para expresar el concepto (M1). Esta presentación en columnas y filas permite observar gráficamente dos de los principales temas de la semántica léxica: la polisemia (en caso de que la misma columna contara con dos entradas, E1,2 - E2,2) y la sinonimia (en caso de que la misma fila contara con al menos dos entradas, E1,1 - E1,2).

En esta matriz léxica los conceptos son representados por la lista de unidades léxicas que pueden ser usadas para expresarlo (todas las entradas que pertenezcan a una misma fila), es decir, el conjunto de sinónimos (*synset*) no explica al concepto sino que simplemente indica que el concepto existe. No obstante, la representación propuesta por esta *matriz léxica* no puede trabajar directamente con los conceptos, sino que lo hace con las unidades léxicas. La relación

léxica principal en WordNet es la sinonimia, pero también están presentes la antonimia, la hiperonimia, la hiponimia, la meronimia y las relaciones morfológicas.

WordNet está organizado en base a relaciones semánticas y como los significados están representados por medio de *synsets* , las relaciones semánticas están representadas como punteros (*pointers*) entre *synsets* . Es decir, a la serie de sinónimos que constituye un *synset* se le suman símbolos que indican el tipo de relación que existe entre un *synset* y otro, por ejemplo @ para indicar los nombres superordinados y ~ para los subordinados de otros *synsets* . Cabe agregar que los tipos de relaciones semánticas varían según la categoría sintáctica en que la que se establezcan.

WordNet, como base de datos léxica, es aplicada a la expansión de búsquedas (*query expansion*) como también a la desambiguación del sentido de las unidades léxicas (*WSD*) en el área de la recuperación y extracción de información.

3.4. Lexicón computacional

'WordNet (WN) ... es tracta del lexicó relacional en format computacional més complet i extens que existeix...'(Martí Antonín 2000: 116)

El interés por el léxico no es nuevo para la Lingüística ni para la Lexicografía como tampoco lo es la elaboración de recursos lingüísticos automatizados o semiautomatizados capaces de manipular el léxico. A partir de las recomendaciones propuestas en la conferencia de 1986 *"Automating the Lexicon: Research and Practice in a Multilingual Environment"* (Walker et al. 1995: 1), diversos grupos de investigación a nivel europeo plantearon proyectos relacionados con la elaboración de lexicones automatizados.

En este sentido el proyecto *EAGLES* (Expert Advisory Group on Language Engineering Standards, 1993-1996) estableció las normas, directrices y recomendaciones para tener en cuenta en la elaboración de lexicones computacionales. El eje central del proyecto *EAGLES* era la reutilización de los recursos lingüísticos con el fin de facilitar el desarrollo de aplicaciones que respondieran a las necesidades reales de los usuarios, e implicaran a su vez, las prácticas, las directrices, las normas y los entornos de trabajo compatibles.

Un lexicón computacional almacena y caracteriza formalmente el conocimiento lingüístico a través de reglas en cada uno de sus niveles de análisis (fonológico, morfológico, sintáctico, semántico y pragmático) que le permiten realizar inferencias. Moreno Ortiz (2000) define el lexicón computacional como *'repositorios de información léxica elaborados con el objeto de servir de soporte representacional a diversas aplicaciones en el ámbito de las tecnologías del lenguaje humano...'*. La finalidad de este recurso es ofrecer información léxica para usuarios no humanos; sus usuarios finales son los sistemas de PLN que adoptan un enfoque basado en conocimiento (*knowledge-based*) que necesitan incorporar conocimiento lingüístico explícito y un conocimiento de carácter general para realizar una tarea específica.

3.5. Tesauro

'WordNet, the on-line English thesaurus and lexical database ...'(Hirst 1999: 628)

La noción de tesauro en Lingüística, tanto en Lexicografía como en PLN, no coincide con su uso en Bibliotecología y Documentación.

El tesauro para la Bibliotecología y la Documentación es considerado un ' *Tipo de lenguaje documental que se integra con términos analizados y normalizados que guardan entre sí relaciones semánticas y funcionales. El tesauro se organiza bajo fuerte control terminológico, con objeto de proporcionar un instrumento idóneo para el almacenamiento y la recuperación de la información en áreas especializadas.* ' (Barité 1997: 145) Es decir que el tesauro es un herramienta documental que se emplea para la indización y recuperación de la información en entornos especializados.

Para la Lingüística un tesauro refiere a un tipo de obra lexicográfica, aunque también implica dos sentidos asociados. El primero considera a la recopilación de las unidades léxicas de una lengua utilizando un criterio diacrónico, pero no selectivo, por ejemplo *'tesoro de la lengua castellana* ' [4] (*Diccionario de la Lengua Española LEMA* 2001: 1739). El segundo considera el tesauro como un repertorio lexicográfico en el que se agrupan unidades léxicas según su significado similar o relacionado y que puede restringirse o no a una lengua determinada, Grefenstette (1994), Kilgarriff (2000, 2003), Curran (2001, 2002).

Un ejemplo clásico de este tipo de obras lexicográficas es el *ROGET's THESAURUS of English Words and Phrases [5]* , cuya finalidad es la de proporcionar ayuda para la búsqueda de los significados de unidades léxicas no frecuentes y seleccionar las formas ortográficas adecuadas. Este recurso no ofrece definiciones para el usuario, con lo cual debe inferir el sentido adecuado de la unidad léxica en cuestión. El tesauro, como recurso auxiliar del discurso tanto escrito como oral, será denominado *tesauro lexicográfico* dado que el uso del término *tesauro* se reserva para la herramienta documental.

3.6. Síntesis

La ontología, la base de conocimiento, la base de datos léxica, el lexicón computacional y el tesauro lexicográfico son recursos lingüísticos cuya finalidad es la *reutilización* de la información que almacenan, e incluso algunos de ellos pueden ser integrados como módulos de un mismo sistema, por ejemplo una base de conocimiento, una ontología y una base de datos léxica.

La divergencia entre estos tipos de recursos lingüísticos deviene del uso impreciso que se realiza en la bibliografía. En gran parte de la bibliografía, las distintas denominaciones de los recursos lingüísticos estudiados son adjudicadas a WordNet, cuya situación y/o interpretación es variable.

Al realizar la revisión bibliográfica también se observaron diversos solapamientos terminológicos entre WordNet y otros recursos lingüísticos. Un primer caso es la confusión con un lexicón computacional (Martí Antonín 2000: 116), recurso con el que WordNet si bien tiene algunos puntos en común (que contiene información léxica y se encuentra en soporte informático), difiere en una de las características esenciales que define a dicho recurso, la presencia de reglas que permitan realizar inferencias. Un segundo caso es el tesauro lexicográfico (Hirst 1999, Kilgarriff 2000), que coincide con WordNet pues incluye unidades léxicas que en determinados contextos pueden reemplazarse unas a otras (sinónimos). Esta confusión quizás sea potenciada por la aplicación WordNet a la expansión de búsquedas (*query expansion*) en el campo de la recuperación de información (Vorhees 1994, Mandala et al. 1999), donde se asume que la función de dicha base de datos léxica es como fuente de sinónimos. Un tercer caso es con la ontología (Berland 1999, Jones 1998, Kietz et al. 2000, Burgun 2001, Fensel et al. 2001). En WordNet, los nombres no están organizados jerárquicamente entorno a un único concepto superordinado tipo entidad que englobe a todos los demás, sino que se agrupan en torno a un conjunto de conceptos denominados *primitivos semánticos* que originan jerarquías léxicas separadas. Esta estructuración plantea el problema del rol de los primitivos, mientras unas veces son tratados como elementos léxicos en otras son tomados como conceptos. Por lo tanto, se plantea la duda de si WordNet trabaja con unidades léxicas o con conceptos. La estructuración por jerarquías léxicas separadas es distinta a la utilizada en las ontologías que inician su jerarquía conceptual a partir un único concepto superordinado (*all*).

Si se comparan entre sí todos los recursos lingüísticos estudiados se pueden realizar dos consideraciones. La primera está en relación al *comportamiento pasivo* o *activo* . Se entiende por *comportamiento activo* cuando el recurso permite realizar inferencias sobre su información, mientras que el *comportamiento pasivo* implica la necesidad de crear aplicaciones específicas para explotar su contenido. Las ontologías, las bases de conocimiento y los lexicones computacionales incluyen reglas que les permiten realizar inferencias sobre la información que contienen. Las bases de datos léxicas y los tesauros lexicográficos son repositorios pasivos, pues si se desea inferir algo sobre su contenido, deben desarrollarse aplicaciones externas para consultar sus datos o deben vincularse con otras herramientas que ya integren mecanismos de inferencia. La segunda consideración, en relación con los elementos sobre los que se organizan estos recursos, supone que las bases de conocimiento y las ontologías se estructuran a través de conceptos, que representan hechos, porciones de la realidad; y las bases de datos léxicas, los lexicones computacionales y los tesauros lexicográficos se estructuran a partir de unidades léxicas.

Particularmente, en el caso de la diferenciación entre base de conocimiento y base de datos léxica, se debería remitir a la distinción entre base de conocimiento y base de datos. Esta diferencia radica en el tipo de contenido que tendrá cada una de ellas, conocimiento y datos respectivamente, con lo cual determinará las características formales y de representación.

4. Los recursos lingüísticos: aplicabilidad desde la documentación

En el apartado 3 se describen las características que singularizan a cada uno de los recursos lingüísticos estudiados. A continuación se estudiará la aplicabilidad y relación de dichos recursos con las tareas documentales.

La ontología es el recurso de aplicación más inmediato para el procesamiento documental ya que proporciona la estructura conceptual de un campo del conocimiento determinado. Por esta razón, es utilizable para la generación de herramientas documentales con fines de indización, búsqueda y recuperación de información.

La base de conocimiento como sistema modular que integre una base textual, una base bibliográfica, una base terminológica y una ontología, es aplicable, tanto para la elaboración de herramientas documentales con fines de indización, búsqueda y recuperación de información, como para la elaboración automática de resúmenes y/o la indización automática. Si se toma como ejemplo la elaboración de un t *esauro enriquecido [6]* , la base textual proporcionaría los contextos de uso, la bibliográfica las fuentes de las cuales fueron tomados esos contextos, la base terminológica la definición y los términos, y la ontología la estructuración del campo del conocimiento en cuestión.

La base de datos léxica y el tesauro lexicográfico se utilizan para la obtención de sinónimos útiles en el campo de la expansión de búsquedas (*query expansion*), y/o también para la elaboración de instrumentos de indización, búsqueda y recuperación de información. Asimismo la base de datos léxica proporciona información que permita seleccionar una forma léxica determinada (nominal, verbal, etc.) que luego sea recuperada en una base terminológica, y a su vez, afecte el diseño de una herramienta documental.

El uso de un lexicón computacional es posible, por ejemplo, para la elaboración automática de resúmenes en el caso que se integre como módulo a un sistema de traducción automática o semiautomática.

Los recursos lingüísticos estudiados tienen una posible reutilización desde el campo de la Documentación, que posibilita una práctica integrada para acercar las perspectivas de trabajo y solucionar las problemáticas comunes.

5. A modo de conclusión

Las tecnologías del lenguaje han proporcionado múltiples recursos lingüísticos, que si bien pueden tener puntos en común, también difieren en su naturaleza y aplicaciones. La utilización de una precisa terminología en la bibliografía especializada contribuiría a comprender mejor las posibilidades de aplicación de cada uno de ellos.

La Lingüística y la Documentación requieren aunar esfuerzos y complementar sus conocimientos para la construcción de herramientas documentales. El concepto de *reutilización* aplicado en la elaboración y uso de los recursos lingüísticos es fundamental en el campo de la Lingüística y debería tenerse en cuenta en la Documentación, con el fin de posibilitar un ahorro de esfuerzos y costes en la generación de herramientas documentales.

Una comprensión de los objetivos, de las prácticas de la Lingüística y de la Documentación posibilitará una labor en común determinada por su objeto de trabajo (la lengua).

Las futuras perspectivas de investigación deberán orientarse a concretar la elaboración de herramientas documentales a partir de la *reutilización* de los recursos lingüísticos, y consolidar así, una práctica tanto interdisciplinar y como transdisciplinar entre la Lingüística y la Documentación.

6. Bibliografía

Alonso Martín, J. A. (2001). La traducció automàtica. En: Martí Antonín, M. A. (coord.). (2001).*Les tecnologies del llenguatge* . Barcelona: Edicions de la Universitat Oberta de Catalunya. 86-119.

Badia Cardús T. (2001). Tècniques de processament del llenguatge. En: Martí Antonín, M. A. (coord.). (2001).*Les tecnologies del llenguatge* . Barcelona: Edicions de la Universitat Oberta de Catalunya. 190-238.

Barité, M. (1997). *Glosario sobre Organización y Representación del Conocimiento. Clasificación, Indización y Terminología* . Montevideo: Escuela Universitaria de Bibliotecología y Ciencias Afines.

Beale, S et al. (1996). *Semantic Analysis in the Mikrokosmos Machine Translation Project* . [en línea]. Baltimore: ILIT. University of Maryland. http://ilit.umbc.edu/SergeiPub/SemantAnalysis.pdf [C onsulta: 29 abril 2004]

Berland, M.; Charniak, E. (1999). *Finding Parts in Very Large Corpora* . [en línea]. Providence: Brown University.

http://citeseer.ist.psu.edu/rd/27382230%2C561441%2C1%2C0.25%2CDownload / http://citeseer.ist.psu.edu/cache/papers/cs/27227/ http:zSzzSzacl.ldc.upenn.eduzSzPzSzP99zSzP991008.pdf/ berland99finding.pdf [Consulta: 29 abril 2004]

Burgun, A.; Bodenreider, O. (2001). *Mapping the UMLS Semantic Network into General Ontologies* . [en línea]. Bethesda, Maryland: National Library of Medicine.

http://citeseer.ist.psu.edu/rd/64207010%2C561993%2C1%2C0.25%2CDownload / http://citeseer.ist.psu.edu/cache/papers/cs/27182/ http:zSzzSzetbsun2.nlm.nih.gov:8000zSzpublis-ob-offizSzpdfzSz2001-amia-ab-Ft.pdf/burgun01mapping.pdf [Consulta: 29 abril 2004]

Cabré, M. T et al. (2004). Base de connaissances GENOMA: le recherche de l'ontologie. En:*Journée d'étude Terminologie, Ontologie & Representation des Connaissances* . Lyon: Université Jean-Moulin Lyon 3. 19-25.

Curran, J. R. (2001). *Automatic Thesaurus Extraction: thesis* . [en línea]. Wivenhoe Park: ICCS, School of Informatics, University of Edinburgh. http://www.cogsci.ed.ac.uk/~jamesc/papers/proposal.pdf [Consulta: 29 abril 2004]

Curran, J. R.; Moens, M. (2001). *Improvements in Automatic Thesaurus* . [en línea]. Wivenhoe Park: ICCS, School of Informatics, University of Edinburgh. http://cswww.essex.ac.uk/staff/poesio/LAC/curran_moens02.pdf [Consulta: 29 abril 2004]

EAGLES (1996). *Welcome to EAGLES on line : Expert Advisory Group on Language Engineering Standards.* [en línea]. http://www.ilc.cnr.it/EAGLES96/home.html [Consulta: 29 abril 2004]

Feliu, J. et al. (2002). *Ontologies: a review* . [en línea]. Barcelona: IULA. Universitat Pompeu Fabra. ftp://ftp.iula.upf.es/pub/publicacions/98inf023.pdf [Consulta: 29 abril 2004]

Fensel, D. et al. (2001). *OIL: An Ontology Infrastructure for the Semantic* . [en línea].

http://citeseer.ist.psu.edu/cache/papers/cs/22554/ http:zSzzSzwww.cs.vu.nlzSz~frankhzSzpostscriptzSzIEEE-IS01.pdf/fensel01oil.pdf/ [Consulta: 29 abril 2004]

Fellbaum, Ch. et al. (1993). *Five papers on WordNet.* Las Cruces: Computing Research Laboratory. [en línea]. http://crl.nmsu.edu/~raz/Ling5801/papers.html

[Consulta: marzo del 2003]

Gamallo Otero, P. (2000 *).* Bases léxicas organizadas mediante un sistema de herencia Mereológica. [en línea]. En*:* *Revista de Procesamiento del Lenguaje Natural* 26. http://www.sepln.org/revistaSEPLN/revista/26/gamallo-otero.pdf [Consulta: 29 abril 2004]

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery* . Boston: Kluwer Academic Publishers.

Guthrie, L.; Pustejovsky, J., Wilks, Y., Slator, Brian M. (1996 *).* The role of lexicons in natural language processing [en línea]. En: *Communications of the ACM* 39(1). 63-72.

http://portal.acm.org/ft_gateway.cfm?id=234204&type=pdf&coll=portal&dl=ACM&CFID=21168780&CFTOKEN=89994111 [Consulta: 29 abril 2004]

Hirst, G. (1999 *).* Review of "EuroWordNet: a multilingual database with lexical semantic networks" by Piek Vossen. Kluwer Academic Publishers 1999 . [en línea]. En: *Computational Linguistics* 25(4). 628-630.

http://delivery.acm.org/10.1145/980000/973236/p628hirst.pdf?key1=973236&key2=1797711801&coll=GUIDE&dl=GUIDE&CFID=19787535&CFTOKEN=83069044 [Consulta: 29 abril 2004]

Johnson, K.; Johnson, H. (eds.). (1998 *). Encyclopedic Dictionary of Applied Linguistics. A Handbook for Language Teaching* . Oxford: Blackwell.

Jones, D. (1998). *Developing Shared Ontologies in Multi-agent Systems* . [en línea]. Liverpool: Department of Computer Science, University of Liverpool.

http://citeseer.ist.psu.edu/rd/35557192%2C112480%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/3259/http:zSzzSzwww.csc.liv. ac.ukzSz%7EdeanzSzpaperszSzECAI98.pdf/jones98developing.pdf [Consulta: 29 abril 2004]

Kietz, J.-U.; Volz, R.; Maedche, A. (2000). Extracting a Domain-Specific Ontology from a Corporate Intranet. [en línea]. En: *Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000* . 167-175.

http://citeseer.ist.psu.edu/cache/papers/cs/27096/http:zSzzSzacl.ldc.upenn.eduzSzWzSzW00zSzW000738.pdf/kietz00extracting.pdf/ [Consulta: 29 abril 2004]

Kilgarriff, A.; Yallop, C. (2000). *What's in a thesaurus* . [en línea]. En: ITRI Reports Series. Brighton: ITRI. University of Brighton.

http://www.itri.bton.ac.uk/techreports/index.html – *ITRI-00-28* [Consulta: 29 abril 2004]

Kilgarriff, A. (2003). *Thesauruses for Natural Language Processing* . [en línea]. En: ITRI Reports Series. Brighton: ITRI. University of Brighton. http://www.itri.bton.ac.uk/techreports/index.html - *ITRI-03-15* [Consulta: 29 abril 2004]

Lema : diccionario de la lengua española. Barcelona : Spes, 2001

Mandala, R.; Tokunaga, T.; Tanaka , H. (1999). Session: Complementing WordNet with Roget's and corpus-based thesauri for information retrieval. [en línea]. En: *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* . Morristown : Association for Computational Linguistics.

http://portal.acm.org/ft_gateway.cfm?id=977049&type=pdf&coll=portal&dl=ACM&CFID=19358554&CFTOKEN=67326162 [Consulta: 29 abril 2004]

Martí Antonín, M. A.; Castellón Masalles, I. (2000 *). Lingüística computacional* . Barcelona: Edicions Universitat de Barcelona.

Mendonça, E. S. (2000). A lingüística e a ciência da informaçao: estudos de uma interseçao. [en línea]. En:*Ciencia da Informaçao* 29(3). 50-70. http://www.ibict.br/cionline/290300/2930006.pdf [Consulta: 29 abril 2004]

Miller, G. (1995). WordNet: a lexical database for english . [en línea]. En: *Communications of the ACM* 38(11). 39-41.

http://delivery.acm.org/10.1145/220000/219748/p39-miller.pdf?key1=219748&key2=9196220801&coll=portal&dl=ACM&CFID=19358554&CFTOKEN=67326162 [Consulta: 29 abril 2004]

Moreiro González, J. A. (1994). Documentación y lingüística: conceptos de relación esenciales. En:*Ciencias de la Información* 25(4). 202-211.

Moreiro González, J. A. (1993). Implicaciones documentales en el procesamiento del lenguaje natural. En:*Ciencias de la Información* 24(1). 48-54.

Moreno Ortiz, A. (2000). Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. [en línea]. En: *Estudios de Lingüística del Español (ELiEs)* 9. http://elies.rediris.es/elies9/ [Consulta: 29 abril 2004]

Mylopoulos, J. (1980). An overview of knowledge representation. [en línea]. En:*Proceedings of the 1980 workshop on Data abstraction, databases and conceptual modeling* 11(16). 5-12.

http://portal.acm.org/ft_gateway.cfmid=806869&type=pdf&coll=portal&dl=ACM&CFID=19358554&CFTOKEN=67326162 [Consulta: 29 abril 2004]

Pérez Hernández, Ch. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. [en línea]. En: *Estudios de Lingüística del Español (ELiEs)* 18. http://elies.rediris.es/elies18/ [Consulta: 29 abril 2004]

Saurí, R et al. (2001). 'Sistemas de representación de la información léxica'. En: Cabré, M. T.; Feliu, J. (eds.). (2001*). La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293)* . Barcelona: UPF. IULA.

Voorhees, E. M. (1994). On expanding query vectors with lexically related words. [en línea]. *En: Proceedings of the 2nd Text Retrieval Conference (TREC-2)*. Gaithersburg: NIST. 223- 231. http://trec.nist.gov/pubs/trec2/papers/txt/21.txt

Walker, D.; Zampolli, A.; Calzolari, N. (eds.). (1995). *Automating the lexicon research and practice in a multilingual* . Oxford: Oxford University Press.

7. Notas

[1] En el presente trabajo unidad léxica se utiliza como sinónimo de palabra

[2] Traducción propia

[3] http://www.cogsci.princeton.edu/~wn/

[4] Este sentido del tesauro en la Lexicografía no se tendrá en cuenta en este trabajo pues no presenta elementos de confusión con los demás recursos que se están estudiando

[5] Versión electrónica en: http://poets.notredame.ac.jp/Roget/

[6] El concepto de tesauro enriquecido refiere a la posibilidad de generar un tesauro con características o informaciones novedosas para este tipo de herramienta documental con la finalidad de proporcionar al usuario final gran cantidad de información sobre un campo temático determinado (contextos de uso, información sobre fuentes consultadas, mayor número de relaciones conceptuales, etc.)