Ontología sobre economía y recuperación de información

Mercè Lorente Casafont

Citación recomendada: Mercè Lorente Casafont. *Ontología sobre economía y recuperación de información* [en linea]. "Hipertext.net", núm. 3, 2005. <http://www.hipertext.net> [Consulta: 30 ene. 2007]. .

1. Introducción

Cuando se habla de recuperación de información (desde ahora RI), podemos referirnos a técnicas distintas para objetivos bien diversos. Los lingüistas, como la mayoría de científicos, somos usuarios privilegiados de muchas de estas técnicas, pero nuestra participación en proyectos de investigación y desarrollo en RI no suele ser tan numerosa como la de documentalistas e informáticos.Las razones hay que buscarlas básicamente en dos constataciones: 1) La mayoría de sistemas de RI tienen por objetivo la recuperación de documentos o la clasificación automática de documentos; y 2) las técnicas matemáticas se han revelado altamente rentables en la ratio de eficacia frente a costes. Todo método independiente de lenguas y de temáticas, si sus resultados en precisión y cobertura *(recall)* son suficientemente positivos, será preferible a otras, sobre todo para operar con grandes volúmenes de datos.

Sin embargo, se ha ido incrementando la incorporación de estrategias lingüísticas en algunos proyectos de RI. El primer motivo reside en qué algunas técnicas de RI han abordado nuevos objetivos, más allá de la recuperación de documentos indizados, como la generación automática de resúmenes, la extracción automática de terminología o la minería de textos. En estos sistemas se busca una información específica dentro de textos o conjuntos de textos, con lo que recobran importancia elementos como el léxico utilizado, la combinatoria léxica y la estructura de estos textos. La segunda causa para la implicación del trabajo lingüístico en proyectos de RI se relaciona con los resultados de estas técnicas.Además de los criterios de evaluación habituales, precisión y cobertura *(recall*), ha adquirido suma importancia el criterio de relevancia. Atribuímos al término relevancia un significado que implica un grado mayor de precisión en la recuperación. El criterio de precisión regula la capacidad del sistema de recuperar documentos pertinentes a la consulta realizada. El concepto de relevancia se refiere sobre todo a la capacidad del sistema por recuperar información adecuada semántica y pragmáticamente dentro de los documentos pertinentes. Por ejemplo, imaginemos que una consulta sobre "minas" da como resultado un 70% de precisión en documentos sobre minería. A partir de aquí se trata de aplicar una medida complementaria para controlar si, entre estos documentos sobre minería, la información que búscábamos es más o menos la que recuperamos. En otras palabras, el criterio de relevancia sería la adaptación del mismo criterio de precisión cuando se aplica a la extracción automática de terminología, sobre los términos que se encuentran dentro de un texto tomados como nodos de conocimiento. Diversas experimentaciones han demostrado que la combinatoria de técnicas estadísticas y lingüísticas permite mejorar ostensiblemente los resultados en precisión en la recuperación de documentos y relevancia en la recuperación de información precisa dentro de estos textos.

En este artículo queremos presentar el proyecto RICOTERM-2 [1] , un proyecto coordinado entre la Universidade de Santiago de Compostela y la Universitat Pompeu Fabra de Barcelona, en el que también participan investigadores de la Universidad del País Vasco y de la Universidad de Amberes. El objetivo principal de este proyecto de investigación, de carácter aplicado, es el diseño y el desarrollo del prototipo de un reelaborador de consultas para búsquedas en Internet (RECCI). Esta aplicación consiste en una interfaz, ubicada en un portal web especializado en economía, que permitirá transformar una consulta simple monolingüe en un conjunto de consultas multilingües expandidas hacia otros conceptos relacionados con la consulta inicial. Las lenguas de trabajo son, además del inglés, el castellano, el catalán, el gallego y el vasco.

En la primera fase del proyecto, nos proponemos construir un banco de conocimiento sobre economía, de carácter modular, compuesto por un corpus textual para cada una de las lenguas de trabajo, un diccionario computacional enriquecido con información semántica y terminológica, y una ontología sobre un ámbito restringido de la economía. El modelo a seguir para la construcción del banco modular es el desarrollado por el grupo IULATERM [2] , en proyectos anteriores, para la genómica, cuya estructura se muestra en la figura siguiente:

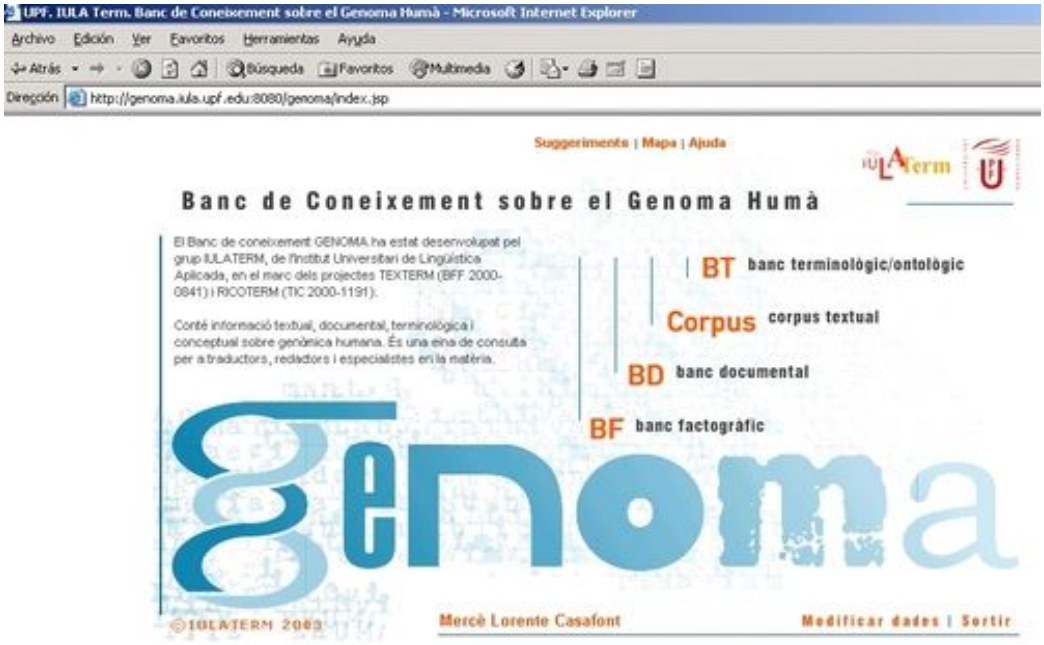

Figura 1 : Pantalla de acceso a los distintos módulos del Banco de Conocimiento sobre Genoma del grupo IULATERM

No obstante, el banco de economía, con un mayor número de lenguas de trabajo, se prevé de dimensiones más reducidas. No contendrá el módulo factográfico y documental y, dado que su utilidad radica en la experimentación del reelaborador, no pretende ser exhaustivo y el contenido de cada uno de los módulos se delimitará cuantitativamente y temáticamente.

Otra característica relevante del banco de conocimiento de economía es que se basa en la idea de reutilización de recursos, de manera que no se están constituyendo corpus textuales ni bases de datos léxicas de nueva planta, sino que se está llegando a acuerdos de cooperación para el uso de materiales ya constituidos en otros proyectos (diccionarios sobre economía, diccionario y corpus en gallego, diccionario y corpus en vasco) y para el uso de herramientas para el procesamiento del gallego y del vasco.

2. La expansión de consultas en RI

Los experimentos en expansión de consultas se basan en métodos que permiten detectar, a partir de la consulta de un usuario, aquello que constituye el núcleo de la consulta (la necesidad informativa del usuario) para expandirla con variantes y encontrar así aquellos documentos que aporten la información más relevante. La primera dificultad con la que se encuentran estos experimentos es precisamente que la consulta no suele ofrecernos suficientes datos para distinguir la información relevante de la relacionada temáticamente (Strzalkowski et al. 1999:136).

Durante los años 90, los métodos de expansión de consultas más usuales han sido la expansión únicamente por términos (*only-term expansion*) y la expansión de texto completo (*full-text expansion*). En el primer caso, la técnica probabilística establece un peso mayor para los términos frecuentes de documentos relevantes, y se suele completar con otros métodos automáticos, como la retroalimentación de relevancia (*relevance feedback method*), o con métodos manuales, como la intervención del usuario en la determinación de la relevancia. En el segundo caso, las consultas se expanden a partir de oraciones o de párrafos, extraídos de documentos considerados relevantes o no.

En los dos casos, el reto más importante es llegar a controlar el contenido semántico y pragmático, que no siempre es explícito a través de las formas de los textos. El procesamiento del lenguaje natural necesita aún bastante desarrollo en representación semántica y pragmática, para evitar que las expansiones partan simplemente de la identificación de unidades léxicas y del establecimiento de correlaciones con variantes morfológicas, equivalentes interlingüísticos o combinaciones léxicas, pero se ha avanzado bastante en los últimos años. En el proyecto RICOTERM-2 nos proponemos abordar el diseño del reelaborador de consultas sobre economía a partir de los dos métodos principales: la expansión a partir de términos y la expansión a partir de textos.

2.1. Expansión de consultas a partir de textos

Los sistemas basados en la expansión de consultas a partir de textos suelen presentar unos resultados de satisfacción sobre un 40% de precisión (Strzalkowski et al. 1999:136), resultados demasiado bajos para usuarios-profesionales que buscan información específica y que tienen expectativas muy distintas a los usuarios comunes que simplemente navegan por la red. Además de la compleción y verificación de la búsqueda a través de la intervención manual del usuario o mediante técnicas automáticas, las soluciones propuestas para mejorar los resultados se basan fundamentalmente en criterios de restricción sobre los textos utilizados en la expansión:

- Restricción estructural: un conjunto estructurado de textos.

- Restricción temática: únicamente documentos de un determinado ámbito.

El proyecto MURAX (Kupiec 1999:314) es un buen ejemplo de restricción estructural aplicada a la expansión de consultas. Se trata de un sistema de pregunta-respuestas (*Question-Answering*), para conseguir respuestas concretas a preguntas concretas del tipo ¿Quién mató al presidente Kennedy? MURAX expande la consulta para localizar informaciones parciales en los textos y combinarlas en una respuesta ideal. El sistema es presentado como una combinación de métodos, que incluye un análisis sintáctico superficial para la extracción de sintagmas. Pero lo que aquí nos interesa es observar que los textos para la expansión provienen de la Enciclopedia Grolier (1990), un texto bien estructurado que, como buen producto lexicográfico, presenta definiciones simples y sistemáticas y controla algunas remisiones léxicas (paráfrasis, sinónimos, hiperónimos). Las experimentaciones del sistema hechas a partir del juego de mesa *Trivial Pursuit* © no obtendrían los mismos resultados si los textos para las extensiones pertenecieran a documentos de estructura muy variada, como contratos, artículos científicos, conferencias, leyes, diálogos, etc.

Consideramos que, en la línea de combinar métodos, la utilización de un corpus seleccionado temáticamente y etiquetado estructural y lingüísticamente podría dar buenos resultados para la expansión de consultas en ámbitos restringidos. La limitación que presentarían estos corpus especializados, como el Corpus Técnico del Institut Universitari de Lingüística Aplicada (http://www.iula.upf.edu), sería su extensión y complejidad, frente a la concisión estructurada de definiciones de un diccionario enciclopédico, como en el proyecto MURAX. No obstante, y siguiendo con la misma dinámica de combinación de técnicas y de recursos, para el proyecto RICOTERM-2 se prevé dos estrategias textuales complementarias para la expansión de consultas:

- El uso de una herramienta de extracción automática de términos (léxico relevante para el ámbito restringido) y el uso de un detector de relaciones conceptuales, que permitan identificarlos y etiquetarlos en contexto.

- Y el uso de bases de datos terminológicas y de diccionarios especializados sobre economía, que incluyan definiciones.

2.2. Expansión de consultas a partir de ontologías

Los sistemas de expansión de consultas que, en los últimos años, están mejorando ostensiblemente sus resultados son los que interactúan con ontologías o jerarquías léxicas. Las ontologías son construcciones formales que representan nodos conceptuales y expresan las relaciones conceptuales que establecen entre sí. Su complejidad y su atomización suele ser mayor que en el caso de los tesauros, ya que su finalidad no es clasificar documentos y localizarlos, sino que se construyen con el fin de ordenar "los conceptos del mundo" [3] y relacionarlos con las expresiones lingüísticas que los vehiculan. Nos referiremos más adelante a los problemas teóricos y aplicados que presentan la construcción de ontologías, pero veamos ahora qué han aportado las ontologías a los sistemas de expansión de consultas.

Cada vez son más las voces que califican la utilización de una ontología para la recuperación de información como un método eficaz, que puede superar a otros métodos en precisión y relevancia. Los sistemas de expansión de consultas que utilizan ontologías, o alternativamente tesauros, se basan en el criterio de expansión de términos o expansión léxica (*only-term expansion*), es decir que a partir del léxico identificado como relevante en un documento se establecen correlaciones con conceptos u otras unidades léxicas que representan estos conceptos o conceptos afines. Desde un punto de vista lingüístico, podríamos decir que a partir de una palabra, un sintagma o un conjunto de palabras de la consulta, el sistema buscaría en la ontología otras palabras o sintagmas que expresaran conceptos próximos. Esta proximidad se correspondería, en lingüística y por regla general, con sinónimos o variantes, con hiperónimos o clases de conceptos, y con cohipónimos o conceptos que pertenecen a una misma clase; difícilmente encontraremos en este tipo de sistemas otro tipo de relaciones conceptuales como las de causalidad o de secuencia temporal.

Uno de los recursos más utilizados como ontología en sistemas de RI con expansión de consultas es WordNet.Esta jerarquía léxica estructura el léxico de las lenguas a partir de la noción de *synset* o conjunto de sinónimos (*synomym sets*), de manera que los sistemas de expansión de consultas asocian automáticamente conjuntos de sinónimos para cada vector de consulta (Voorhees 1994: 223). Uno de los primeros problemas detectados en este tipo de sistemas es que, al tratarse de un recurso de carácter general, no restringido temáticamente, las consultas simples suelen estar sujetas a ambigüedad (polisemia para los lingüistas). Esta limitación se suele resolver en experimentación mediante la intervención manual del usuario que selecciona el *synset* relevante (Voorhees 1994: 223) y mediante técnicas de *machine learning* .

Las alternativas a ontologías generalistas como WordNet, que mejoren resultados en precisión y sobre todo en relevancia, pasan por la delimitación del alcance de la ontología o estructura léxica. Hay casos de ontologías generadas *ad hoc* a partir de las palabras y de los conceptos usados en el seno de una corporación (empresa, organismo público, red de trabajo), experimentaciones que trabajan con léxico controlado y con pocos datos, y que difícilmente pueden ser exportables a otras corporaciones, a otras lenguas o a la RI abierta en Internet (Stenmark 2003: 9).

Otros proyectos, en cambio, delimitan su alcance a un ámbito especializado, generalmente con una tradición consolidada en la comunidad científica, y por lo tanto con un consenso muy amplio sobre conceptos, relaciones conceptuales y terminología estandarizada. Es el caso ejemplar de los trabajos de RI en medicina, que ilustramos aquí con la referencia de un proyecto de expansión de consultas por asociación a conceptos del UMLS Metathesaurus© (Aronson et al. 1997: 485). El UMLS Metathesaurus © forma parte de los recursos desarrollados por la *National Library of Medicine* , contiene información sobre conceptos y términos biomédicos en diversas lenguas, y se construye automáticamente a partir de vocabularios controlados y de sistemas de clasificación (Feliu et al. 2002: 24). En su versión del 2001 su capacidad era de 800.000 conceptos y de 1.400.000 lemas o palabras distintas. Podemos prever, en este caso, que los buenos resultados de la experimentación del sistema a partir de un conjunto de textos seleccionados, algunos de los cuales estaban indizados y otros no, podrían ser similares con otro conjunto de textos especializados seleccionados y una ontología completa de otro ámbito de especialidad. De hecho en este proyecto, se concluye que la combinación de los métodos de expansión de consultas a partir del UMLS Metathesaurus © y a partir de citas de MEDLINE © , que incluyen tanto técnicas probabilísticas como técnicas lingüísticas, pronostican unos resultados aún mejores.

De acuerdo con el criterio de reutilización de recursos existentes, en el proyecto RICOTERM-2 se está trabajando en la localización de estructuraciones conceptuales sobre economía que puedan ser incorporadas, por accesibilidad y compatibilidad de formatos, a una ontología base. Hemos podido constatar también que la presencia de conceptos de este ámbito restringido es muy alta en WordNet, frente a otros campos de especialidad menos representados. Otro criterio que guía la construcción de la ontología en RICOTERM-2 es precisamente la adquisición automática de conceptos y de relaciones conceptuales a partir de textos reales. En este sentido, resulta primordial la utilización de un extractor automático de términos y un detector de relaciones conceptuales, a los que hacíamos referencia en el apartado anterior.

3. La construcción de ontologías

No quisiéramos dejar de lado una cuestión que se plantea a menudo en este tipo de proyectos de lingüística aplicada. Por un lado, es innegable que la experimentación ha demostrado que el uso de ontologías, sobre todo en el caso de ontologías de dominios restringidos o especializados, permite mejorar en precisión y en relevancia tanto las consultas como las respuestas de los sistemas de RI. Sin embargo, a menudo se reflexiona sobre la adecuación teórica y descriptiva de este tipo de recursos. Vamos a ver algunas de las ventajas y de los problemas que suscitan.

3.1. Ventajas de las ontologías de ámbito especializado

Las ontologías de ámbitos especializados presentan una granularidad muy elevada. Se representan en ellas clases de conceptos, conceptos nodulares y relaciones conceptuales diversas, resultado de los procesos de conceptualización y de categorización realizados en un campo de especialidad. En esto superan sin lugar a duda a los tesauros documentales, en los que generalmente predominan las relaciones jerárquicas y asociativas, y en los que la función primordial es relacionar conceptos con categorías determinadas previamente. Al menos en el plano teórico, las ontologías permiten aproximarnos a los conceptos de una materia con un planteamiento más abierto, con muchas más relaciones conceptuales, y no están restringidas *a priori* por una finalidad práctica. Para mostrar la granularidad, véase a continuación un ejemplo de la ontología del Genoma Humano a la que nos referíamos anteriormente:

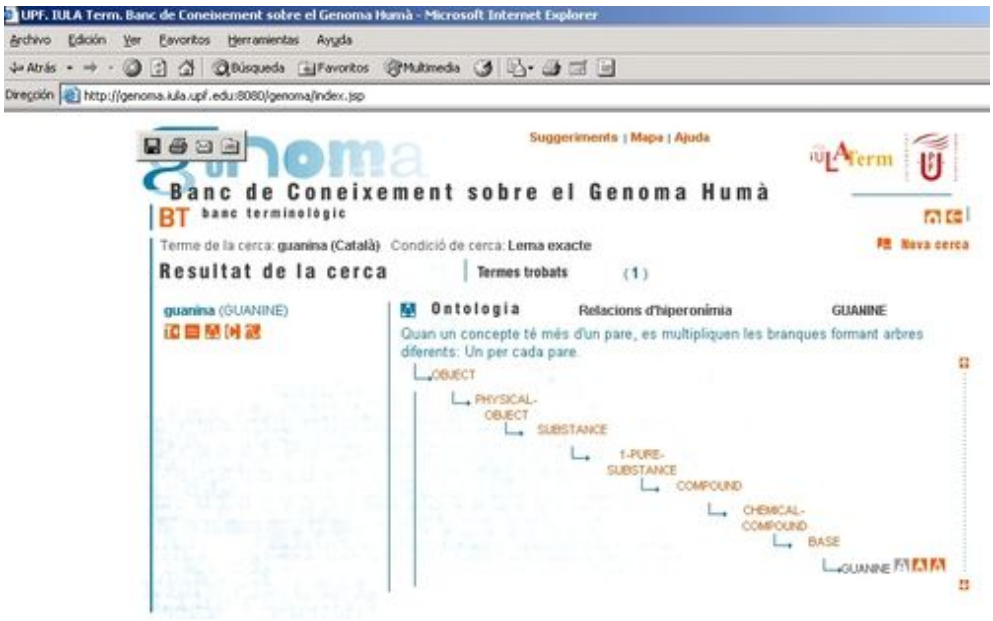

Figura 2 : Pantalla de visualización de la ontología (ejemplo de relaciones de hiperonímia de guanina)

Las ontologías de ámbitos especializados de larga tradición académica y profesional (y no es casual que expresemos esta limitación) muestran una gran estabilidad . Disciplinas como la medicina o la biología han experimentado cambios, desarrollos e innovaciones muy drásticos, pero esto no significa que estas novedades invaliden completamente organizaciones conceptuales anteriores. De manera que una ontología sobre patologías o sobre farmacología podría verse afectada por el avance científico (ubicación de conceptos, propiedades asignadas a conceptos, creación de nuevos conceptos, fragmentación de conceptos), pero difícilmente será reformada en su totalidad. La causa de esta estabilización en la ontología de una disciplina, así como la estabilización en la terminología utilizada, es el alto nivel de consenso internacional que muestran algunas de estas disciplinas, favorecido por la ágil transferencia de conocimiento y por el control que ejercen organismos de estandarización.

Las fuentes utilizadas para la construcción de ontologías de ámbitos especializados suelen ser ejemplos de comunicación controlada . Tanto si se trata de artículos científicos, enciclopedias y diccionarios, manuales académicos, documentación de profesionales, conferencias y ponencias científicas, documentación técnica, como si se trata del mismo conocimiento experto de los usuarios especialistas, los contenidos y su expresión suelen adecuarse al registro discursivo formal. Además los textos escritos, y los orales que han sido previamente elaborados para ser leídos o semiimprovisados, son sometidos a diversos procesos de revisión por sus autores o por mediadores lingüísticos (traductores, redactores, correctores).

En síntesis, la granularidad de la representación conceptual, la estabilidad de estos conceptos en la comunidad científica y la formalidad de la gran mayoría de documentos que transmiten estos conceptos y sus relaciones convierten las ontologías de los ámbitos especializados en un recurso fiable para la recuperación de información, y más efectivo que sus correlatos de ámbito no restringido y para el lenguaje común.

3.2. Problemas de las ontologías

Claro que no debemos esperar que, con una ontología de ámbito especializado, se nos resuelvan todos los problemas en la recuperación de información específica. Los resultados en índices de satisfacción de los usuarios pueden ir mejorando con el desarrollo de nuevas ontologías específicas y, sobre todo, con el uso de técnicas combinadas, pero no debemos ignorar que el mismo desarrollo de ontologías presenta problemas de fondo, a los que nos quisiéramos referir brevemente en este apartado a partir de dos contradicciones o paradojas.

1. La paradoja lingüística. Desde una perspectiva lingüística, los conceptos no existen con independencia de la denominación o del conjunto de denominaciones que los vehiculan. De hecho, la única acción verdaderamente onomasiológica [4] es la creación de términos por parte de los especialistas de una materia (científicos que descubren un elemento, tecnólogos que construyen un artefacto, profesionales que quieren distinguir un elemento parcialmente nuevo de otros existentes, etc.) y esto siempre se hace desde una lengua. Además la terminología actual (Cabré 1999: 149) asume que los términos, como todos los signos del lenguaje, no son independientes de las lenguas y que la comunicación especializada está condicionada por aspectos sociales y culturales. En este sentido una de las primeras paradojas con la que debemos enfrentarnos los constructores de ontologías que trabajamos en terminología y compartimos esta visión sociolingüística es precisamente el uso de gestores o interfaces que utilizan el inglés como si se tratara de una lengua franca a la que podemos asociar sin

problemas los equivalentes del resto de lenguas de trabajo, o lo que es peor como si fuera simplemente la representación neutral de una estructuración conceptual común. Obviamente la lingüística aplicada, con su carácter eminentemente pragmático, no dejará de construir ontologías por esta contradicción, sino que buscará alternativas metodológicas que satisfagan más a los usuarios y a los mismos constructores, como el uso de caracteres alfanuméricos o la consideración de la propia ontología como una jerarquía léxica [5] autónoma para cada lengua de trabajo.

2. El dinamismo de los conceptos . La contradicción que se establece entre la elaboración de una ontología como una representación estable y la consideración de que los conceptos y las estructuraciones de conceptos son entidades dinámicas ha sido puesta de manifiesto tanto por filósofos como por lingüistas y terminólogos. Incluso en medicina informática y en algunos foros sobre economía (Vromen, 2004:213) hemos podido constatar que existe un debate abierto entre los partidarios del uso de ontologías para la representación de la información y los críticos, que ven las limitaciones de las ontologías para representar conceptos de la economía evolucionista por ejemplo. Para los cambios que impone el paso del tiempo en la revisión de conceptos y la creación de otros nuevos, la lingüística aplicada tan solo puede ofrecer una defensa continuada de la necesidad de actualizar este tipo de recursos permanentemente; pero esta solución tiene un obstáculo que suele ser insalvable: los costes asociados. En cambio, para la representación en ontologías de conceptos dinámicos,no estáticos, las propuestas aún se encuentran en la fase de primeras tentativas (hipervínculos, tratamiento de la polisemia y de la ambigüedad, objetos difusos, etc.) y deberemos esperar un tiempo para analizar propuestas concretas con ejemplos reales.

4. Primeros diseños del sistema RECCI

Como introducíamos al inicio, el proyecto RICOTERM-2 se propone como objetivo principal el diseño y la elaboración de un prototipo para el sistema RECCI (Reelaborador de Consultes per a Cercadors d'Internet). La primera versión del prototipo hemos querido implementarla en relación con la recuperación de información en economía, porqué esta temática ya dispone de recursos textuales y léxicos especializados, porqué estamos en disposición de compilar corpus textuales en cada una de las lenguas de trabajo del proyecto (castellano, catalán, gallego, vasco e inglés), y porqué consideramos que se trata de un tema prototípico para la necesidad de información actualizada por parte de los usuarios de sistemas en recuperación de información. Los agentes que podrían estar interesados en un sistema de este tipo, ubicado en un portal especializado, podrían ser colegios profesionales, medios de comunicación, administraciones públicas, agentes de bolsa, bancos y cajas de ahorro, y otras empresas y colectivos profesionales.

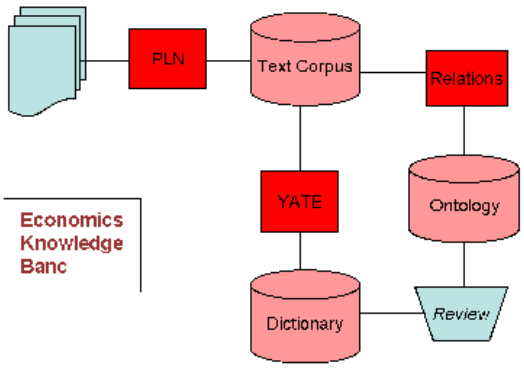

Figura 3 : Arquitectura del banco de conocimiento sobre economía del proyecto RICOTERM-2

En la figura anterior se puede observar la arquitectura del banco de conocimiento sobre economía, previsto como campo de pruebas para el sistema RECCI. Los corpus textuales de cada una de las lenguas de trabajo son corpus etiquetados estructuralmente y procesados lingüísticamente, de manera que podremos extraer de ellos información por unidades léxicas y por categorías gramaticales, y por secuencias más amplias y por patrones sintácticos. Aunque se ha previsto reutilizar estructuraciones conceptuales y diccionarios de economía ya existentes, no descartamos como parte de la construcción del banco de conocimiento explotar los corpus textuales etiquetados para la extracción automática y semiautomática de términos y de relaciones conceptuales entre términos.

Por este motivo, uno de los objetivos secundarios del proyecto consiste en la adaptación para la economía y las lenguas de trabajo citadas de la herramienta YATE, un extractor automático de candidatos a términos independiente de lenguas

y que combina estrategias probabilísticas y lingüísticas (Vivaldi 2001). Como ampliación de las estrategias lingüísticas de esta herramienta, se prevé la integración de un diccionario de expresiones lingüísticas que vehiculan relaciones conceptuales (Feliu 2004).

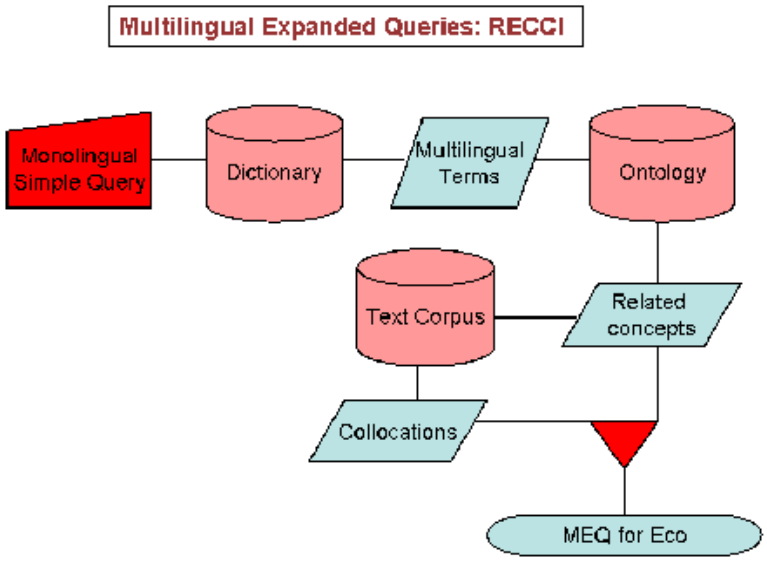

Figura 4 : Diseño de la expansión de consultas del proyecto RICOTERM-2

En esta figura podemos observar el diseño inicial del proceso de expansión de consultas. Partimos de la idea de que los usuarios, aunque sean profesionales, suelen realizar consultas simples en una única lengua. El sistema prevé que esta consulta simple monolingüe sea ampliada, mediante la consulta del diccionario, con términos equivalentes en las otras lenguas y con variantes sinonímicas. Esta ampliación de la consulta multilingüe y variada, como entrada a la ontología, nos permitiría detectar también otros términos asociados y, como entrada a la herramienta de consulta de corpus, nos permitiría detectar también colocaciones o combinatoria lingüística en contexto.

A modo de conclusión, el diseño del sistema RECCI se corresponde con una expansión de consultas multilingüe que combina técnicas distintas. Además de métodos matemáticos a los que no nos hemos referido en este artículo, las técnicas lingüísticas utilizadas serán tanto la expansión de base léxica como la expansión mediante corpus. Consideramos que la utilización de un banco de conocimiento modular para el ámbito de la economía, que incluya textos procesados lingüísticamente, diccionarios y una ontología, nos permitirá que el reelaborador de consultas ofrezca unos buenos resultados, consistentes en consultas complejas y multilingües, que reflejen preguntas ideales de los profesionales para lanzar a motores de búsqueda usuales.

El proyecto continuará en fases posteriores con experimentaciones del sistema con datos reales del banco de conocimiento de economía y con evaluaciones del grado de satisfacción por parte de un colectivo piloto de usuarios, constituido por documentalistas y economistas.

5. Síntesis

En este artículo, además de presentar los primeros pasos del proyecto RICOTERM-2, hemos querido dar respuesta a algunas de las inquietudes que se plantean en la toma de decisiones en este tipo de proyectos de carácter aplicado.

Hemos expuesto que se trata de un proyecto que pretende aunar esfuerzos en el desarrollo de recursos lingüísticos para ámbitos especializados y en el diseño de sistemas de recuperación de información que mejoren las prestaciones para con los usuarios y que amplien sus expectativas sobre la precisión de la información recuperada. En esta línea, reivindicamos tres aspectos fundamentales del proyecto aplicado: la reutilización de recursos existentes, la orientación hacia ámbitos especializados de uso no restringido, el multilingüismo integrador de las lenguas oficiales en el estado español y la adecuación del diseño de recursos a perfiles de usuarios reales (no ideales).

Para completar la presentación del proyecto, hemos introducido los primeros esquemas sobre la arquitectura del banco de conocimiento sobre economía y sus componentes modulares y hemos presentado un primer esbozo sobre el funcionamiento de la expansión de consultas multilingüe.

Por otro lado hemos aprovechado la ocasión para revisar algunas cuestiones necesarias para abordar este diseño: desde

la tipología de expansiones de consultas existentes a la reflexión sobre la adecuación del uso de ontologías en un marco teórico de la terminología de base comunicativa.

6. Bibliografía

Angelova, Galia (2000) "Ontologies for Natural Language Processing Applications". En: Simov, Viril; Kiryakov, Atanas (eds.) (2000) *Ontologies and Lexical Knowledge.* Sofia: OntoText Lab, Sirma AI EOOD, p. 1-15.

Aronson, Alan R.; Rindflesch, Thomas C. (1997) "Query Expansion Using the UMLS ® Metathesaurus ® ". En: *Journal of the American Medical Informatics Association* , 1997: Supplement. Proceedings of the 1997 AMIA Annual Fall Symposium, p. 485-489.

Buitelaar, Paul (2000) "Semantic Lexicons: Between Terminology and Ontology". En: Simov, Viril; Kiryakov, Atanas (eds.) (2000) *Ontologies and Lexical Knowledge.* Sofia: OntoText Lab, Sirma AI EOOD , p. 16-24.

Cabré, M. Teresa (dir.) (2004) Banc de coneixement sobre Genoma Humà. IULATERM © . Consultado: 28-feb-2005, http://genoma.iula.upf.edu:8080/genoma/index.jsp

Cabré, M. Teresa (1999) *La terminología: representación y comunicación.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. ISBN 84-477-0673-7.

Feliu, Judit (2004) *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra , 2004. ISBN: 84-89782-19-9. [CD-ROM]

Feliu, Judit; Vivaldi, Jorge; Cabré, M. Teresa (2002) *Ontologies: a review.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra , Papers de l'IULA, Sèrie Informes, 34.

Kupiec, Julian M. (1999) "MURAX: Finging and Organizing Answers from Text Search". En: Strzalkowski, Tomek (ed.) (1999) *Natur al Language Information Retrieval.* Dordrecht: Kluwer Academic Publishers, p. 311-332. ISBN 0-7923-5685-3.

Lorente, Mercè (2005) "Expansió de consultes multilingüe per a la recuperació d'informació en economia". En:*Actas del XXIII Congreso de AESLA* (Mallorca, marzo 2005). En prensa.

Madsen, Bodil Nistrup; Pedersen, Bolette Sandford; Thomsen, Hanne Erdman (2000) "Semantic Relations in Content-based Querying Systems. A Research Presentation from tge OntoQuery Project. En: Simov, Viril; Kiryakov, Atanas (eds.) (2000) *Ontologies and Lexical Knowledge.* Sofia: OntoText Lab, Sirma AI EOOD , p. 72-81.

Simov, Viril; Kiryakov, Atanas (eds.) (2000) *Ontologies and Lexical Knowledge Bases. 1 st Internacional Workshop, Ontolex 2000 Sozopol, Bulgaria, September 8-10, 2000; Proceedings*. Sofia: OntoText Lab, Sirma AI EOOD.

Stenmark, Dick (2003) "Query Expansion Using an Intranet-based Semantic Net". En: Proceedings of IRIS-26. Porvoo: Information Systems Research in Scandinavia Association. http://w3.informatik.gu.se/~dixi/publ/qe-iris.pdf

Strzalkowski, Tomek(ed.) (1999) *Natural Language Information Retrieval.* Dordrecht: Kluwer Academic Publishers. ISBN 0-7923-5685-3.

Strzalkowski, Tomek; Lin, Fang; Wang, Jin; Pérez-Carballo, José (1999) "Evaluating Natural Language Processing Techniques in Information Retrieval". En: Strzalkowski, Tomek(ed.) (1999)*Natural Language Information Retrieval.* Dordrecht: Kluwer Academic Publishers, p. 113-145. ISBN 0-7923-5685-3.

Vivaldi, Jorge (2001) *Extracción de candidatos a término mediante combinación de estrategias heterogéneas* . Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, 2004, Sèrie Tesis 9.ISBN 84-89782-11-3. [CD-ROM]

Voorhees, Ellen M. (1994) "On Expanding Query Vectors with Lexicaly Related Words". En: Harman, D.K. (ed.) NIST Special Publication 500-215: The Second Text Retrieval Conference (TREC-2). Gaithersburg: NITS, p. 223-232.

Vromen , Jack (2004) " Conjectural revisionary economic ontology: Outline of an ambitious research agenda for evolutionary economics". En: *Journal of Economic Methodology* , 2004, vol. 11:2, p. 213-247

7. Notas

[1] Este artículo se inscribe en el marco del proyecto RICOTERM-2. Control terminológico y discursivo para la recuperación de información en ámbitos comunicativos especializados, mediante recursos lingüísticos específicos y un reelaborador de consultas (HUM2004-05658-C02-01), financiado por el Ministerio de Educación y Ciencia, Plan Nacional de I+D+I (2004-07). [volver]

[2] IULATERM es un grupo de investigación consolidado del Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, dedicado a la investigación básica y aplicada en léxico, terminología, discurso especializado, gestión del conocimiento e ingeniería lingüística, formado actualmente por 19 doctores y otros tantos colaboradores y

dirigido por la Dra. M. Teresa Cabré Castellví. [volver]

[3] Ponemos entre comillas esta expresión de los conceptos del mundo por tratarse de un objeto inalcanzable, dinámico, y siempre mediatizado cultural y lingüísticamente. [volver]

[4] Partimos de un concepto que delimitamos en nuestra mente y le asignamos una expresión lingüística para identificarlo, representarlo y comunicar algo sobre él. [volver]

[5] Como el caso de WordNet y EuroWordNet, ontologías creadas desde una aproximación lingüística. Sin embargo no debemos olvidar que el proceso de trabajo en EuroWordNet ha sido bastante dependiente de WordNet, elaborado para el inglés de USA. [volver]