SOCIEDAD ESPAÑOLA PARA EL
PROCESAMIENTO DE LENGUAJE NATURAL

Técnicas lingüísticas aplicadas a la búsqueda textual multilingüe

Ambigüedad, variación terminológica y multilingüismo

Anselmo Peñas Padilla

Resumen

Los sistemas de búsqueda han adquirido una gran importancia en el uso cotidiano de los ordenadores. Sin embargo, la recuperación de información textual tiene asociada una serie de problemas todavía no resueltos satisfactoriamente. Algunos de estos problemas provienen de las características propias del lenguaje natural. Por esta razón, diversos autores se han interesado en la aplicación de técnicas lingüísticas automáticas a la recuperación de información, obteniendo resultados que hasta la fecha no son plenamente satisfactorios y que cuestionan la utilidad de estas técnicas en la búsqueda textual. En este trabajo se abordan los problemas de ambigüedad léxica, variación terminológica y translingüismo en el acceso a la información, con la siguiente línea argumental:

1. Estudio del papel de las técnicas lingüísticas en el modelo tradicional de recuperación y ordenación de documentos.
2. Transición a un modelo interactivo en el que los resultados parciales del procesamiento lingüístico se ofrecen al usuario como caminos alternativos de contextualización de la consulta y de acceso a la información.
3. Creación de un marco en el que sea posible la evaluación de estos sistemas interactivos de acceso a la información.

La primera parte muestra una serie de experimentos de recuperación con el fin de discernir si la falta de buenos resultados se debe a los errores que introduce el procesamiento automático o si se debe a que las técnicas lingüísticas no resultan estrategias adecuadas en un modelo tradicional de recuperación de documentos. Estos experimentos se han llevado a cabo sobre una colección etiquetada manualmente en todos los niveles léxicos. De esta forma, los resultados de recuperación quedan libres de los errores de un procesamiento automático permitiendo determinar si las técnicas lingüísticas (en una situación ideal) suponen o no estrategias adecuadas para mejorar la recuperación. Los experimentos llevados a cabo muestran que, en un modelo tradicional de recuperación de documentos, ni la desambiguación de la categoría gramatical, ni la detección y distinción de compuestos léxicos, ni la desambiguación del sentido de las palabras producen mejoras significativas en la recuperación de documentos que justifiquen el coste de

procesamiento que introducen. La desambiguación del sentido de las palabras, sin embargo, permite realizar una indexación basada en *synsets* de WordNet que resulta prometedora y que abre la posibilidad de una recuperación translingüe gracias al índice interlingua de EuroWordNet. Los experimentos que se presentan muestran que esta indexación conceptual basada en *synsets* resulta bastante robusta ante la introducción de errores. Estos resultados llevaron a la decisión de implementar un prototipo que permitiera evaluar no sólo cuantitativamente, sino también cualitativamente una recuperación multilingüe basada en indexación conceptual sobre synsets de EuroWordNet. La evaluación de este sistema arroja a la luz varios retos por resolver que de momento no hacen efectiva la recuperación basada en indexación conceptual.

Con estos antecedentes, la segunda parte del trabajo explora una nueva posibilidad de abordar los problemas de ambigüedad léxica, variación terminológica y multilingüismo. En lugar de aplicar las técnicas lingüísticas a la indexación de información, subordinándolas al modelo clásico de recuperación de documentos, estás técnicas se utilizan para ofrecer al usuario un nivel de procesamiento lingüístico parcial en un modelo interactivo de acceso a la información. Este nuevo nivel de información se ha concretado, en este trabajo, en una nueva área de terminología extraída automáticamente a partir de la colección y particularizada de acuerdo con la consulta. La indexación de información se dirige a obtener y normalizar los sintagmas de la colección. En el momento de la consulta, cuando el usuario realmente puede expresar sus necesidades de información, el sistema le ayuda a contextualizar su consulta sugiriéndole sintagmas presentes en la colección. Estos sintagmas suponen variaciones morfosintácticas, semánticas y translingües de su consulta y, a la vez, vías de acceso directo a los documentos. Este modelo da forma a la interfaz del sistema *Website Term Browser (WTB)* en la que aparece, además del área tradicional de documentos, un área de terminología recuperada, seleccionada y organizada de acuerdo con la consulta.

Las evaluaciones diseñadas para los sistemas de recuperación de documentos no son aplicables a *WTB.* Tampoco resultan apropiadas las evaluaciones diseñadas para los sistemas de búsqueda interactiva, ni para los sistemas que proporcionan interactividad al tratar los problemas de multilingüismo. Por esta razón, el tercer punto en torno al cual se estructura este trabajo es la creación de un nuevo marco de evaluación para sistemas interactivos con las características de *WTB*. La evaluación muestra que los usuarios estiman de utilidad el nuevo nivel de información terminológica, sirviendo de complemento al ranking tradicional de documentos. La evaluación también muestra la capacidad del sistema para recuperar información multilingüe y para tratar grandes volúmenes de información.

Abstract

Information Retrieval systems play an important role in the daily use of computers. Nevertheless, there are still a number of problems related to Text Information Retrieval that have not as yet been resolved satisfactorily. The origin of some of these problems come from the natural language characteristics. For this reason various authors have attempted to the apply automatic linguistic techniques to Information Retrieval. However, the results obtained so far are not completely satisfactory and challenge the usefulness of these techniques in document retrieval. In the present work we deal with problems of lexical ambiguity, terminological variance and multilinguality in Information Access, given the following structure:

1. Study of the role of linguistic techniques in traditional document retrieval and ranking model.
2. Transition to an interactive model in which intermediate results of the linguistic processing are presented to the user to suggest alternative ways of query contextualization and information access.
3. Development of an evaluation framework for these interactive information access systems.

The first part of this work, presents a series of document retrieval experiments. The objective is to find out whether the lack of good results is due to errors introduced by automatic processing or is due to the inadequateness of linguistic techniques in the traditional document retrieval model. These experiments have been designed and carried out using a collection manually annotated at all lexical levels. In this way, the results are free from automatic processing errors, which permits us to determine whether linguistic techniques offer strategies to improve text retrieval. The experiments carried out show that in a traditional document retrieval model, neither grammar category disambiguation, lexical compound detection and distinction, nor sense disambiguation yield significant improvements in document retrieval. However, sense disambiguation allows indexing to be undertaken based on WordNet *synsets,* which opens up the possibility of cross-language information

retrieval using the EuroWordNet (EWN) InterLingual Index. The experiments presented here show that conceptual indexing based on *synsets* is quite robust against possible errors. Based upon these results we decided to implement a prototype that would not only allow quantitative but also qualititative evaluation of cross-language information retrieval based on conceptual indexing using EWN *synsets*. The evaluation showed a number of challenges that need to be resolved, leading to the conclusion that, at the moment, information retrieval based on conceptual indexing will not improve performance results obtained with traditional models.

Given this background, the second part of this work explores a new possibility of tackling the problems of lexical ambiguity, terminological variance and multilinguality. Instead of applying linguistic techniques to document indexing, subordinating them to the classical information retrieval model, these techniques are used to offer the user the results of intermediate linguistic processing in an interactive information access model. This new level of information is materialized in a new area of terminology which is automatically extracted from the collection and characterized according to the query. The purpose of information indexing is to obtain and standardize the phrases in the collection. At querying time, at the moment when users express their information needs, the system helps them to contextualize the query offering phrases taken from the collection. These phrases are morphosintactic, semantic or cross-language variations of the query, providing direct access to documents. This model underlies the Website Term Browser (WTB) interface in which, apart from the traditional document area, there is a new area where the retrieved terminology is selected and ranked in according to the query.

Current evaluation tests designed for document retrieval systems are not suitable for WTB. Neither are the evaluation techniques designed for interactive retrieval systems or for systems that provide interactivity dealing with multilinguality. This situation motivates the third main part of this work: the creation of a new evaluation framework for interactive systems with the WTB characteristics. The evaluation performed in this work shows that users consider the new level of terminological information useful, as it complements the traditional document ranking outcome. The evaluation also shows the system's capability to retrieve cross-language information and to deal with large volumes of data.

Índice

Tablas

Figuras

Capítulo 1

Introducción

Los sistemas de búsqueda han adquirido una gran importancia en el uso cotidiano de los ordenadores hasta el punto de que realizar una consulta en un buscador es la acción más frecuente, tras el envío de un correo electrónico. Sin embargo, la búsqueda y recuperación de información textual tienen asociadas una serie de problemas todavía no resueltos satisfactoriamente. Algunos de estos problemas provienen de la ambigüedad y la falta de estructura propias del lenguaje natural.

1.1 Barreras del lenguaje en Recuperación de Información

Considérese una situación de búsqueda en la que un profesor quiere acceder a *recursos* sobre *educación especial* que le ayuden en el aula. Podría emitir la siguiente consulta en un buscador:

recurso educación especial

El primer documento que recupera uno de los mejores buscadores en Internet es una orden ministerial que contiene:

"... recurso contencioso / administrativo ... educación especial ..."

1.1.1 Ambigüedad léxica

El ejemplo muestra el problema de la ambigüedad del lenguaje natural en recuperación de información: la palabra *recurso* es polisémica, como ocurre con la mayoría de las palabras más frecuentes.

El ejemplo muestra también la limitación de los interfaces de búsqueda para expresar objetivos de búsqueda. En la mayoría de los buscadores, esta limitación obliga a que finalmente los usuarios recorran y filtren personalmente los documentos recuperados por el sistema. Resultaría deseable obtener un nivel intermedio de información que facilitara esta labor.

Además de los problemas originados por el carácter inherentemente ambiguo del lenguaje natural existen otros problemas que también afectan a la recuperación de información. Por ejemplo, la consulta anterior no recupera documentos que contienen expresiones como:

1. *recursos educativos*
2. *medios audiovisuales en la enseñanza*
3. *special needs education*

1.1.2 Variación morfosintáctica

La primera expresión evidencia el problema de variación morfosintáctica. Una consulta que contiene "*recursos de educación"* no puede recuperar un documento por contener "*recursos educativos"* a pesar de que sean expresiones equivalentes. En este caso la variación es fruto de un cambio de categoría gramatical a través de un proceso de morfología derivativa. Esta variación morfosintáctica es una *permutación*, pero también pueden darse *inserciones* (v.g. *recursos audiovisuales de educación* sería una variación de *recursos de educación*), o variaciones por *coordinación* (v.g. *recursos culturales y educativos* sería una variación de *recursos educativos*).

1.1.3 Variación semántica

La segunda expresión del ejemplo, *medios audiovisuales en la enseñanza*, además de una variación morfosintácticca de *inserción*, es una variación semántica puesto que contiene sinónimos de las palabras originales (*recursos de educación*). En los buscadores tradicionales, la palabra *recurso* no puede proporcionar acceso a los documentos que contienen su sinónimo *medio*, ni la palabra *educación* puede

recuperar documentos con el sinónimo *enseñanza*, a pesar de que, conceptualmente y a efectos de búsqueda, sean equivalentes.

La aproximación más evidente para abordar este problema es la expansión de la consulta añadiendo a la misma palabras sinónimas a las originalmente presentes. Sin embargo, la expansión automática de la consulta con palabras sinónimas no hace más que mostrar la dificultad de la tarea. Por ejemplo, *educación* tiene como sinónimos palabras tan diversas como *ademán, alimentación, capacitación, civilidad, crianza, cría, enseñanza, forma, formación, instrucción, manera, modales, preparación, etc.* Como es de esperar, la introducción de estos sinónimos en la consulta produce resultados, cuanto menos, impredecibles.

1.1.4 Variación translingüe

La tercera expresión del ejemplo muestra el mismo problema en el caso translingüe[1]. La consulta "*educación especial*" no puede recuperar un documento que contenga "*special needs education*" aunque contenga traducciones directas de las palabras originales de la consulta. Una vez más, la traducción automática de la consulta *palabra a palabra* no resulta una solución viable. Por ejemplo, *educación* puede traducirse por *breeding, civility, education, fostering, instruction, manners, nurture, nurturing, pedagogy, politeness, raising, rearing, teaching, training, upbringing, etc.* Su inclusión sin más en la consulta distorsiona gravemente los resultados de la búsqueda.

Expansión y traducción muestran, de nuevo, el problema de la ambigüedad del lenguaje natural. No sólo las palabras a traducir tienen varios sentidos, sino también las palabras que sirven de traducción.

[1] El concepto de acceso translingüe a la información (del inglés cross-language) se refiere al caso en el que una consulta en un idioma permite recuperar información en un idioma diferente. El concepto de acceso translingüe a la información es más específico que el concepto de acceso multilingüe, mereciendo la pena aclarar su distinción. El acceso multilingüe a la información incluye a los sistemas que consideran información en varios idiomas pero únicamente acceden a información en el mismo idioma que la consulta. Un sistema que permite recuperar información en un idioma diferente al idioma de consulta proporciona una funcionalidad adicional y nos referiremos a él como un sistema translingüe de acceso a la información. A lo largo del trabajo utilizaremos indistintamente los términos de *recuperación multilingüe* y *recuperación translingüe* para referirnos a este último.

1.2 Situaciones de búsqueda limitadas por las barreras del lenguaje

La primera cuestión que debe responderse es si realmente estos problemas afectan a la recuperación de información y en qué situaciones se presentan. La respuesta depende del perfil de los usuarios, del tipo de necesidad de información y del grado de precisión de los objetivos de búsqueda.

1.2.1 Presupuestos de los modelos estándar de Recuperación de Información

Los modelos tradicionales de Recuperación de Información se basan en la búsqueda y ordenación, por relevancia decreciente, de una lista de documentos que se ajustan a la consulta. Detrás de este planteamiento hay una serie de presupuestos que no siempre se cumplen en todas las situaciones de búsqueda:

1. El objetivo es recuperar todos los documentos relevantes y sólo ellos, es decir, maximizar simultáneamente la precisión y la cobertura.
2. Las necesidades de información permanecen estáticas, independientemente de los documentos explorados por el usuario.
3. El valor se encuentra en el conjunto de documentos obtenidos.

Sin embargo, la realidad es que en muchas situaciones de búsqueda la consulta varía continuamente, los usuarios se mueven por una variedad de fuentes, y nuevas informaciones conducen a nuevas ideas y nuevas direcciones de búsqueda. Es decir, los presupuestos implícitos en el modelo tradicional de búsqueda obvian la interacción del usuario con el sistema. En estas situaciones, el valor no se encuentra en el conjunto de documentos recuperados, sino que el valor de la búsqueda se encuentra en los fragmentos recogidos a lo largo del proceso de búsqueda (Bates 1990).

1.2.2 Situaciones de imprecisión

En una situación de imprecisión el usuario no sabe o no puede expresar de forma concreta el objeto de su búsqueda. En estos casos el usuario sigue una estrategia que no siempre resulta efectiva: iniciar el proceso con una consulta general y poco a poco refinarla a partir de los resultados proporcionados por el sistema. Esto, en los sistemas actuales, debe realizarse sin ningún tipo de asistencia y requiere la exploración de un gran número de documentos.

Aún así, hay dos situaciones en las que esta estrategia tampoco va a proporcionar buenos resultados:

1. El usuario no conocen la colección[2], no es experto del dominio y no conoce la terminología propia del mismo.

2. El usuario tiene un vocabulario pasivo en varios idiomas pero activo sólo en el idioma propio. Es decir, aunque puede entender textos en otros idiomas, únicamente se expresa con suficiente corrección en el idioma propio.

En estos casos, los problemas de ambigüedad, variación terminológica y variación idiomática dificultan la consecución de los objetivos de búsqueda. Se trata de barreras que el usuario se ve obligado a romper por sí mismo sin ayuda del sistema. Sería deseable, por tanto, que los niveles intermedios de información que deben ayudar al usuario a expresar y concretar sus objetivos de búsqueda, ayuden también a superar estas barreras del lenguaje.

1.3 Técnicas automáticas, terminología y acceso a la información

Abordar estos problemas requiere, cuanto menos, la consideración de técnicas, recursos y herramientas de procesamiento lingüístico, es decir, estudiar cómo el Procesamiento de Lenguaje Natural (NLP) puede ayudar a superar las barreras lingüísticas en Recuperación de Información (IR).

Sin embargo, puesto que la recuperación de información se realiza sobre los términos que contienen los textos, los problemas de ambigüedad, variación terminológica y translingüismo también se han tratado desde el área de Recuperación de Información evitando el uso de técnicas NLP.

Es conveniente, por tanto, visualizar el papel que juega la Terminología[3] como un área independiente de las áreas de Recuperación de Información y de Procesamiento de Lenguaje Natural, pero con evidente intersección con ambas.

---

[2] A lo largo del trabajo se entenderá por *colección (de documentos)* al conjunto de textos que se han procesado e indexado y sobre los que se realiza el proceso de recuperación de información. Es una concepción estática de la colección en la que la adición de un nuevo documento implica la creación de una nueva colección y acarrea un nuevo procesamiento.

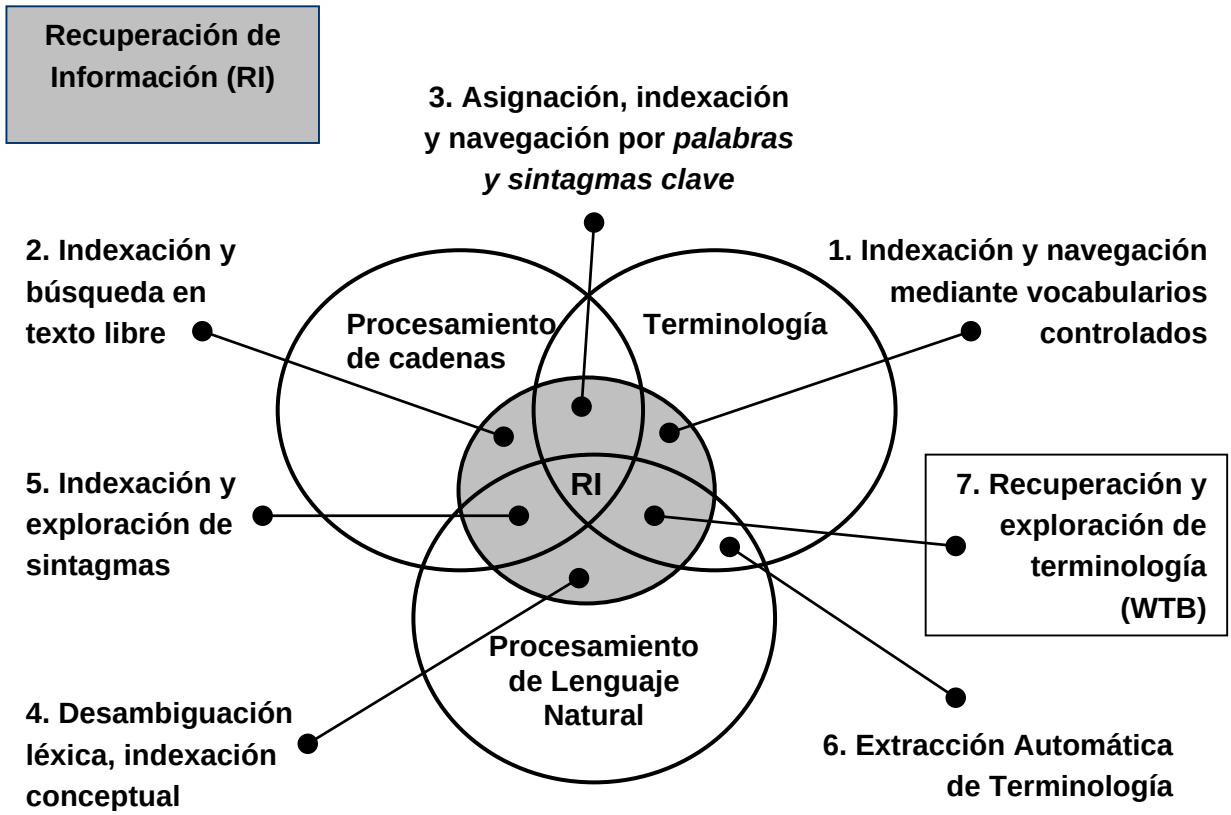


Figura 1-1. Intersecciones NLP, SP y Terminología con respecto a IR

Las intersecciones entre estas cuatro áreas (NLP, SP, Terminología e IR) permiten situar el problema, así como las perspectivas para abordarlo.

1. *IR – Terminología*. La intersección entre IR y el área de Terminología viene representada por el uso de vocabularios y tesauros en la recuperación de información, tanto en la indexación de documentos con los términos del vocabulario, como en la exploración y navegación del vocabulario para acceder a los documentos de la colección. La construcción y mantenimiento de vocabularios controlados, así como la asignación de términos del vocabulario a cada uno de los documentos, suponen tareas con un coste muy elevado que motiva la aparición de aproximaciones automáticas.

[3] *Terminología* se utiliza aquí no con el sentido de *vocabulario* en un dominio específico, sino como *conjunto de prácticas y métodos dirigidos a la recolección, descripción y presentación de términos*. *Término* se utilizaría así como vocablo o expresión propia de un determinado *vocabulario*, ya sea controlado (un tesauro), ya sea una *lista terminológica* extraída automáticamente, ya sea el conjunto de términos de indexación de una colección en IR.

2. *IR - Procesamiento de Cadenas Alfanuméricas (String Processing, SP).* La búsqueda en texto libre puede verse como la confluencia de SP en IR. Las técnicas de SP se utilizan, junto con información estadística, en la identificación y pesado de términos (fundamentalmente palabras) de acuerdo con su valor para discriminar los textos en que aparecen. La búsqueda en texto libre es una aproximación que evita el uso de vocabularios controlados externos al contenido en sí de los documentos. De esta forma no es necesario construir, mantener los vocabularios controlados, ni asignarlos a los documentos.

3. *IR – SP – Terminología.* Entre las dos aproximaciones anteriores aparece la posibilidad intermedia de seguir utilizando vocabularios controlados pero tratando de asignar automáticamente a los documentos los términos de indexación apropiados pertenecientes al vocabulario. Con este punto de partida aparecen sistemas que identifican los sintagmas que podrían conformar un supuesto vocabulario de la colección, permitiendo la navegación entre documentos a traves de los mismos.

4. *IR - NLP*. Debido a que la ambigüedad léxica es uno de los problemas que afectan a la recuperación, una de las intersecciones entre NLP e IR es el empleo de herramientas de desambiguación léxica (categoría gramatical y sentido) para desambiguar los términos de indexación. Otra intersección importante viene determinada por los esfuerzos de indexación conceptual, ya sea por utilizar normalizaciones de expresiones lingüísticas complejas, ya sea por el empleo de redes semánticas y ontologías. Por último, también se puede inscribir en esta intersección el uso de recursos léxicos para la expansión y traducción de consultas y documentos.

5. *IR - NLP – SP*. En los últimos años, los sintagmas extraídos estadísticamente han empezado a utilizarse también como medio de navegación y exploración de la colección (*Phrase Browsing*). La posibilidad de que las técnicas lingüísticas ayuden a superar las limitaciones de estos sistemas ha despertado el interés mutuo entre las comunidades de *IR y NLP* que convergen en este área (Wacholder 2001).

6. *NLP – Terminología*. La actividad más representativa de la intersección entre el área de NLP y el área de Terminología es la de Extracción Automática de Terminología (TE), si bien puede haber otras tareas cercanas como el reconocimiento de Entidades Nombradas. NLP proporciona técnicas para facilitar la creación de vocabularios y recursos léxicos que, a su vez, también se utilizan en algunos procesamientos lingüísticos.

7. *IR – NLP – Terminología*. A partir de las metodologías propias de Extracción Automática de Terminología va a ser posible desarrollar una aproximación de recuperación y exploración de terminología. Esta es la aproximación que se va a presentar en la segunda parte de esta monografía, y que va a permitir acceder a la información considerando variaciones terminológicas morfosintácticas, semánticas y translingües.

1.4 Objetivos

Las técnicas y herramientas lingüísticas que pueden ayudar a superar los problemas de ambigüedad, variación terminológica y translingüismo son las relativas a desambiguación léxica, tanto de la categoría gramatical como del sentido de las palabra, y las herramientas que permiten considerar unidades lingüísticas complejas como colocaciones y sintagmas nominales.

A pesar del número de tareas en las que NLP resulta potencialmente útil para la recuperación de información, los intentos por aplicar a la recuperación de información las técnicas de NLP en cualquiera de los niveles lingüísticos no han tenido mucho éxito (Strzalkowski 1999) (p. xvii). Una de las causas puede encontrarse en que las técnicas NLP no hayan alcanzado un nivel suficiente de desarrollo. La otra explicación posible es que las técnicas NLP no se hayan aplicado a las tareas adecuadas de acceso a la información. Estas dos explicaciones determinan los objetivos de este trabajo, que son:

1. Abordar los problemas de ambigüedad léxica, variación morfosintáctica, variación semántica y variación translingüe en el acceso a la información.

2. Determinar el papel que pueden desempeñar, a este respecto, las técnicas lingüísticas automáticas en los modelos tradicionales de recuperación y ranking de documentos. Realizar una serie de experimentos de recuperación de documentos cuyos resultados queden libres de errores de procesamiento automático y permitan determinar si las técnicas lingüísticas pueden suponer o no estrategias adecuadas para mejorar la recuperación.

3. Desarrollar nuevos marcos de aplicación de las técnicas lingüísticas en la tarea de acceso a la información que permitan abordar los problemas de variación terminológica y multilingüismo.

4. Desarrollar un sistema que demuestre la viabilidad de la propuesta resolviendo los posibles problemas técnicos.

5. Diseñar un marco de evaluación adecuado a las características del sistema desarrollado.

1.5 Estructura del trabajo

En el Capítulo 2 se exponen algunos conceptos básicos así como los trabajos relacionados en el área. Tras este capítulo, la memoria se estructura en dos partes. La primera (Capítulo 3) trata de determinar el papel de las técnicas lingüísticas en un modelo tradicional de recuperación y ranking de documentos. La segunda parte (Capítulos 4, 5 y 6) encuentran en la *exploración de sintagmas terminológicos* un buen punto de partida para aprovechar las técnicas lingüísticas en un nivel de información más próximo a la interacción del usuario y así abordar los problemas de variación morfosintáctica, semántica y translingüe. Finalmente, el Capítulo 7 recoge las conclusiones y plantea futuras líneas de trabajo.

1.5.1 Parte I: ambigüedad léxica e indexación conceptual

La primera parte del trabajo (Capítulo 3) consta de una serie de experimentos que tratan de esclarecer si las causas de que las técnicas lingüísticas no puedan mejorar significativamente la recuperación deben encontrarse en las estrategias utilizadas o en las perturbaciones inherentes a la falta de precisión en todo procesamiento automático.

La metodología de estos experimentos es la siguiente: definición del experimento sobre la base de las hipótesis a evaluar, realización del experimento y análisis de resultados.

Los resultados de los experimentos realizados apoyan la viabilidad de un sistema basado en indexación conceptual con suficiente margen de error en las técnicas automáticas de procesamiento de lenguaje natural como para desarrollar un motor de búsqueda con este modelo. Esto motivó la realización de un trabajo colectivo (*ITEM Search Engine*) que se presenta brevemente, dirigido a implementar el modelo basado en indexación conceptual y evaluar cualitativamente su viabilidad.

Sin embargo, la experiencia del motor de búsqueda *ITEM* no resultó satisfactoria debido, fundamentalmente, a tres factores: el excesivo coste de procesamiento, la

falta de precisión en las técnicas de desambiguación del sentido de las palabras, y a que los conceptos de la red semántica se corresponden con unidades léxicas demasiado pequeñas para la traducción.

Estos resultados apoyan la tesis ya apuntada en el estado de la cuestión de que resulta muy difícil que las técnicas lingüísticas aplicadas únicamente a la indexación pueden mejorar significativamente la recuperación de información dentro del esquema clásico de los motores de búsqueda. Es necesario replantear el concepto de acceso a la información para encontrar nuevos paradigmas que se vean enriquecidos por las técnicas de procesamiento lingüístico[4].

1.5.2 Parte II: acceso interactivo a la información mediante exploración de sintagmas

La segunda parte del trabajo (Capítulo 4, 5 y 6) explora la posibilidad de aplicar las técnicas lingüísticas de bajo coste computacional a la exploración de sintagmas como vía de acceso a la información. La consideración de sintagmas es, por si misma, una manera de abordar la ambigüedad léxica. El sintagma es un contexto capaz, en gran medida, de desambiguar implícitamente el sentido de las palabras componentes del sintagma. Sin embargo, los sistemas actuales de exploración de sintagmas no permiten abordar los problemas de variación morfosintáctica, semántica y translingüe. Para abordar estos problemas, la segunda parte de la monografía propone un paradigma interactivo en el que el sistema realiza un procesamiento del lenguaje acorde con la tecnología actual, y el usuario realiza las elecciones finales que implican la comprensión real de los objetivos de búsqueda. El procesamiento del lenguaje se dirige a extraer la terminología (palabras y sintagmas) de la colección de forma automática para que el sistema la ofrezca al usuario como un nivel intermedio de acceso a la información, anterior a los documentos. Estos sintagmas incluyen variaciones morfosintácticas, semánticas y translingües de la consulta que previamente han sido identificadas en la colección.

---

[4] Los sistemas de búsqueda de respuestas (QA, *Question Answering*) son un ejemplo de intento de superación de los modelos tradicionales de IR. La entrada a un sistema de QA es una pregunta formulada correctamente en lenguaje natural a la cual debe responder el sistema con un párrafo que contenga la respuesta a la pregunta. Estos sistemas suelen incluir motores de búsqueda pero además tienen que abordar los problemas de representación del conocimiento, modelado de preguntas, etc. Tal como se está definida la tarea de QA, ésta se centra en búsquedas precisas cuya respuesta se encuentra en un parrafo. Este escenario no se ajusta al escenario planteado en este trabajo, se trata una tarea muy amplia en la que los problemas de ambigüedad, variación terminológica y translingüismo no son centrales.

De esta manera, el sistema ofrece al usuario dos áreas, una de terminología y otra de documentos. La relación entre ambas áreas permite al usuario interactuar con el sistema para determinar la información que busca aunque no venga expresada ni en la forma ni el idioma de la consulta. El modelo se implementa y evalúa en el sistema Website Term Browser (WTB).

Las carencias de algunos sistemas actuales de exploración de sintagmas se deben a que no usan procesamiento ni recursos lingüísticos por el coste computacional que acarrean. Sin embargo, el modelo propuesto muestra que la tarea de exploración de sintagmas permite relajar el procesamiento lingüístico automático haciéndolo viable desde un punto de vista computacional sin perder las ventajas que aporta el uso de estas técnicas y recursos lingüísticos: la consideración de variaciones morfosintácticas, semánticas y translingües.

El modelo propuesto basado en interactividad sobre términos (palabras y sintagmas) extraídos automáticamente de la colección requiere el desarrollo de nuevas formas de evaluación. Las experiencias en el apartado interactivo de las conferencias TREC se realizan en condiciones de laboratorio, sobre consultas pre-establecidas y con usuarios controlados, que no se ajustan a las necesidades de evaluación del modelo propuesto. La última parte del trabajo muestra cómo se ha evaluado, en un entorno real de trabajo cuyos usuarios tienen necesidades reales de información, la utilidad de los sintagmas en el acceso interactivo y translingüe a la información. También muestra la capacidad del sistema para recuperar terminología de forma translingüe, y su capacidad para tratar grandes cantidades de información.[Φ]

[Φ] Este trabajo ha sido financiado por los siguientes proyectos:
EuroWordNet (CE, IV Programa Marco, LE#4003)
ETB: European Schools Treasury Browser (CE, V Programa Marco, IST-1999-11781)
ITEM: Recuperación de Información Textual en un Entorno Multilingüe (CICyT TIC96-1243-C03-01)
RILE: Servidor de Recursos para el Desarrollo de la Ingeniería Lingüística en Español (ATYCA TS41/99)

Capítulo 2

Preliminares

Tras la introducción de algunos conceptos básicos de Recuperación de Información y Procesamiento del Lenguaje Natural, el capítulo discute, en primer lugar, los trabajos que conducen a la afirmación de que la indexación con técnicas lingüísticas, no es satisfactoria. En segundo lugar, se discutirán los trabajos que tratan de discernir el papel que juega la ambigüedad léxica en la recuperación de información.

La falta de buenos resultados conduce a la conclusión de que las técnicas lingüísticas deben aplicarse a otras tareas de más alto nivel en el acceso a la información: interpretación y contextualización de la consulta, concreción y refinamiento del objeto de búsqueda, presentación de niveles intermedios de información, etc. Es decir, en las tareas que permiten ofrecer información al usuario sobre la que interactuar con el sistema. En este trabajo se aborda esta perspectiva desde la exploración de terminología extraída de forma automática a partir de la colección. Por tanto, la tercera parte del estado de la cuestión se dirige a repasar los diferentes métodos, técnicas y sistemas de exploración de términos con fines de acceso a la información.

Las conclusiones recogen las limitaciones de todas estas aproximaciones sobre todo en cuanto a la consideración de variaciones morfosintácticas, semánticas y translingües en recuperación de información.

2.1 Conceptos básicos

2.1.1 Recuperación de Información

Recuperación de Información (IR, Information Retrieval) es el término utilizado para referirse a la tarea de proporcionar información sobre la existencia de documentos relevantes para una petición de información. El conocimiento del área está suficientemente estructurado en libros como (Frakes 1992), (Baeza-Yates 1999;Baeza-Yates 1999) o (Witten 1999b) por lo que aquí se van a presentar de forma breve los conceptos más relevantes para este trabajo.

La Recuperación de Información, en su acepción tradicional, se dirige idealmente a obtener todos y solo aquellos documentos relevantes para una consulta en una colección dada de documentos. La mayoría de los sistemas devuelven un ranking de los documentos de acuerdo a su grado de relevancia respecto a la consulta.

La tarea de Recuperación de Información se divide en tres subtareas:

1. Indexación, cuya finalidad es obtener una representación de los documentos que permita, de forma eficiente, su almacenamiento y comparación con las consultas.
2. Procesamiento de la consulta, para adecuarla a la representación interna de los documentos. En la mayoría de los motores de búsqueda actuales existe un lenguaje intermedio con operadores que permiten pesar y relacionar entre sí las palabras de la consulta.
3. Recuperación y ranking de documentos a partir de la comparación entre las representaciones internas de consulta y documentos.

También se puede entender como subtarea IR, no siempre presente en todos los buscadores, la posibilidad de retroalimentación (feedback) por parte de los usuarios al sistema. La selección de uno o varios documentos permite al sistema utilizar esta información para refinar la consulta (relevance feedback). En ocasiones el sistema asume que los primeros documentos del ranking son más relevantes y los utiliza directamente para refinar los resultados sin interacción con el usuario (pseudo-relevance feedback).

El objeto de la indexación es caracterizar los documentos para la tarea de recuperación. Idealmente, la indexación de un documento debería ser una representación eficiente del contenido semántico del documento original.

Al igual que los índices indican en un libro en qué pagina aparece una palabra, los índices en Recuperación de Información permiten localizar un término en un documento de una colección. Los términos de indexación pueden ser de dos tipos:

1. Términos de un vocabulario controlado (tesauros). En este caso se asignan al documento una serie de términos pertenecientes a un vocabulario que no tienen por qué estar contenidos en el documento. La asignación puede realizarse de forma manual, semiautomática o automática mediante sistemas entrenados.

2. Texto libre. La indexación de un documento se realiza con los términos más significativos extraídos automáticamente del documento. Estos términos usualmente son palabras del texto, pero nada impide que un procesamiento apropiado sea capaz de extraer elementos más complejos como nombres propios, sintagmas, expresiones normalizadas para fechas y números, incluso la categoría o el sentido de las palabras.

Los tesauros llevan utilizándose mucho tiempo como vocabularios controlados respecto a los cuales indexar y recuperar documentos (Bernier 1957) (Vickery 1960). Cuando se hizo posible trabajar con bases documentales extensas y considerar los textos completos, se comenzaron a desarrollar métodos de indexación automáticos basados en procesamiento lingüístico y estadístico (Salton 1971) sobre texto libre. Estas técnicas de búsqueda libre proporcionaron unos resultados que cuestionaron el uso de los tesauros para recuperación de información por el coste que supone construir un tesauro, mantenerlo y asignar sus términos a los documentos de la colección.

2.1.1.1 Indexación mediante los términos de un tesauro

Un tesauro es un vocabulario controlado, estructurado en relaciones jerárquicas y asociativas, destinado generalmente a almacenar y recuperar los documentos de un sistema de información determinado, recurriendo al uso de palabras clave para referirse a su contenido.

En general, las entradas de un tesauro contienen:

- **Descriptores**, que son los términos que se utilizarán obligatoriamente para representar los conceptos en la indexación de los documentos y en la formulación de las preguntas.
- **No-descriptores**, que son sinónimos o términos que designan conceptos muy próximos a los representados por los descriptores, a la vez que constituyen

puntos de acceso en el tesauro. Su finalidad es lograr que la terminología de la indexación y recuperación converjan en los términos preferidos, es decir, los descriptores.

- **Notas de aplicación**, que explican sucintamente el sentido en que se utiliza el descriptor cuando éste es ambiguo.
- **Relaciones de equivalencia semántica** entre descriptores y no-descriptores.
- **Relaciones de jerarquía** entre descriptores genéricos (BT, Broader Term) y específicos (NT, Narrower Term).
- **Relaciones de asociación**, de ideas entre descriptores relacionados (RT, Related Term).
- **Relaciones de equivalencia lingüística** entre descriptores que designen el mismo concepto en diferentes idiomas.

Los tesauros permiten estructurar y clasificar información, presentarla de acuerdo con modelos de datos comunes (plantillas de indexación) y acceder a los recursos mediante la navegación por la estructura del tesauro. Esto permite no sólo facilitar la publicación, sino también la recuperación de recursos. Los tesauros, además, pueden ser multilingües y esto permite paliar en alguna medida el problema de acceder a documentos de idiomas diferentes.

Sin embargo, el problema fundamental de utilizar tesauros es el coste que tiene asociado su construcción y mantenimiento, así como la asignación de los términos adecuados de clasificación a los documentos. Además, por tratarse de vocabularios controlados en el que cada uno de los términos (descriptores) utilizados sólo tiene un significado, el dominio de aplicación tiene que ser suficientemente específico. Una nueva colección en un nuevo dominio requiere la construcción de un nuevo tesauro.

2.1.1.2 Indexación de texto libre

El foco de interés de este trabajo no se centra en la asignación a los documentos de términos de un vocabulario controlado (keyword assessment), sino a la indexación a partir de términos extraídos automáticamente del contenido de los textos.

Aunque estos términos de indexación pueden ser expresiones más complejas, es habitual que un término de indexación se reduzca a una única palabra. En este caso, se suelen transformar las palabras para reducir el vocabulario a considerar en la indexación. El procesamiento más común consta de tres pasos:

1. Convertir todas las letras a minúsculas (*case folding*). Esto permite, por ejemplo, que las palabras que empiezan por mayúscula tras un punto, así

como las palabras en mayúsculas de rótulos y títulos sean consideradas de igual manera. Como contrapartida, pueden confundirse acrónimos y nombres propios, perdiendo esta información.

2. Reducción a una forma base, raíz o stem. Por ejemplo, *recuperar*, *recuperación, recuperaciones, recuperable*, etc. podrían reducirse a una forma común como *recuper o recup*. En el caso de lenguas poco flexivas como el inglés, la reducción de una palabra a una forma base puede realizarse mediante algoritmos sencillos de truncamiento de sufijos (*stemming)*. Estos algoritmos, que dan buenos resultados si la morfología es muy sencilla como en el caso (dominante) del inglés, no funcionan tan bien en lenguas más flexivas como es el caso del español. En estos casos puede resultar conveniente la utilización de analizadores morfológicos.

3. Omisión de *stop words* o palabras vacías de contenido, palabras como preposiciones, determinantes, etc., que por aparecer en casi todos los documentos no tienen valor discriminativo.

Es necesario notar que la indexación únicamente mediante palabras del texto no permite abordar los problemas de polisemia y sinonimia, ya que considera las palabras sin distinguir sus posibles significaciones.

En general, los términos extraídos de un documento se pesan de acuerdo con sus frecuencias de aparición como estimación de su capacidad para identificar el documento que los contiene:

1. *Term frequency, tf(t)*: número de ocurrencias del término t en el documento. Un término es más representativo de un documento cuantas más veces aparezca en él.

2. *Document frequency, df(t):* número de documentos indexados por el término t, es decir, número de documentos en la colección que lo contienen. Cuanto más raro sea el término (menor *df*), mayor representatividad tendrá en los documentos que lo contienen. Por esta razón suele utilizarse su inversa (*Inverse document frequency, idf*).

Las medidas clásicas que integran estos dos valores para *pesar* o cuantificar la representatividad de un término *t* en un documento conforman lo que se denomina familia *tf.idf(t)* como, por ejemplo:

$$tf(t)\cdot log(N/df(t))$$

donde N es el número de documentos en la colección.

Existen varios métodos de indexación, es decir, métodos que permiten localizar un término en un texto: ficheros invertidos (*inverted files*), ficheros de firma (*signature files*), mapas de bits (*bitmaps)*, etc. El más sencillo es el fichero invertido cuyas variantes también son las que proporcionan mejores resultados en casi todos las situaciones de búsqueda (Witten 1999b).

La siguiente tabla muestra un ejemplo de fichero invertido. A cada término se le asocia la lista de documentos en los que aparece. Esta información puede enriquecerse con el peso del término en el documento, las posiciones dentro del documento, el número de segmento u oración en que aparecen, etc. Toda esta información será utilizada para seleccionar y ordenar los documentos que se ajusten a la consulta.

Número	Término	(Documento; posiciones)
1	universidad	(1;2,14) (4;8)
2	distancia	(1;6) (3;9)
3	lenguaje	(2;3)
4	natural	(2;4) (9;5,16, 27)
…	…	…

Por ejemplo, el término *universidad* aparece en dos documentos, el 1 y el 4. En el documento 1 ocupa las posiciones 2 y 14; en el documento 4 ocupa la posición 8. Para saber en qué documentos aparece un término basta con ir a la correspondiente entrada de la tabla. Para buscar un documento que contenga varios términos se toman sus respectivas entradas en la tabla y se obtienen la intersección de los documentos que aparecen en las listas. Para recuperar los documentos que contienen un determinado sintagma se hace lo mismo pero observando que además de aparecer en el mismo documento, los términos aparecen en posiciones contiguas.

2.1.1.3 Evaluación en IR

Los criterios de evaluación más extendidos en Recuperación de Información son la *precisión* y la *cobertura o recall*[5] *(Figura 2-1).* Por *precisión* se entiende la fracción de documentos relevantes recuperados entre la totalidad de los documentos recuperados por el sistema. Por *cobertura* se entiende la fracción de documentos

[5] A lo largo de este trabajo se utilizará indistintamente *cobertura*, el vocablo inglés *recall* o la expresión *índice de recuperación*.

relevantes recuperados entre la totalidad de documentos relevantes que contiene la colección para una determinada consulta.

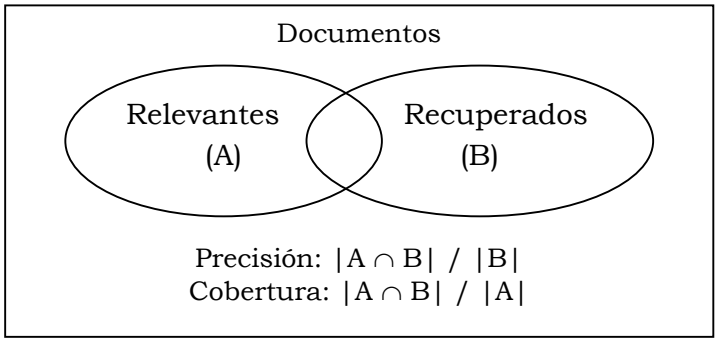


Figura 2-1. Precisión y cobertura en IR

La *precisión* trata de capturar la capacidad de un sistema para recuperar únicamente documentos relevantes mientras que la medida de *cobertura (recall)* trata de capturar la capacidad del sistema para recuperar la *totalidad* de los documentos relevantes. Ambas medidas están sujetas a un juicio subjetivo de relevancia de los documentos para cada consulta del test.

Para considerar varias consultas y comparar varios sistemas se utilizan las curvas de *precisión / cobertura (Salton 1983a)*. En estas curvas se dan valores medios de *precisión* calculados para niveles fijos de *cobertura*. Estos valores se obtienen por interpolación a partir de las posiciones, en el ranking devuelto por el sistema, de todos los documentos relevantes. La forma de proceder es la siguiente:

1. *Obtención de los puntos de observación*. Supongamos que a la consulta Q_1 se le han asignado 4 documentos relevantes y que el sistema S_1 los ha recuperado en las posiciones 1, 2, 10 y 25. Entonces se tienen los siguientes *puntos de observación* [*cobertura, precisión*]: [1/4, 1/1], [2/4, 2/2], [3/4, 3/10] y [4/4, 4/25].

2. *Interpolación*. Si se quiere conocer cuál es la *precisión* para un valor arbitrario de *cobertura R* se realiza una interpolación de orden 0 hacia a la izquierda, es decir, se toma el valor de *precisión* del primer *punto de observación* a la derecha de R. En el ejemplo, para conocer el valor de *precisión* para una *cobertura* R=0.4, se tiene que el primer *punto de observación* con *recall* mayor o igual que R es [0.5, 1] por lo que la *precisión* que se obtiene es 1.

3. *Media en los puntos fijos de cobertura*. La gráfica *precisión / cobertura* más utilizada es la de *precisión media* en 11 puntos de *cobertura* correspondientes a 0, 0.1, 0.2, ..., 0.9 y 1. La forma de construir esta gráfica es calcular para

cada consulta los valores de *precisión* en los 11 puntos de *cobertura* y hacer la media de todas las consultas en cada punto. Por ejemplo, los valores medios de *precisión* en los 11 puntos de *cobertura* correspondientes a 3 consultas Q_1, Q_2 y Q_3, en un sistema S_1 serían:

Cobertura	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Q_1	1	1	1	1	1	1	0.3	0.3	0.3	0.3	0.16
Q_2	0.8	0.6	0.5	0.45	0.45	0.3	0.3	0.25	0.2	0.2	0.1
Q_3	1	0.7	0.6	0.4	0.3	0.2	0.2	0.1	0.1	0.1	0.1
Precisión media S_1	0.93	0.77	0.7	0.62	0.58	0.5	0.27	0.22	0.2	0.17	0.12

4. *Gráfica comparativa*. La gráfica obtenida permite comparar el comportamiento de varios sistemas diferentes sobre un mismo test. Por ejemplo, a continuación se muestra la gráfica para una tabla de resultados de tres sistemas sobre el mismo test:

Cobertura	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
S_1	0.93	0.77	0.7	0.62	0.58	0.5	0.27	0.22	0.2	0.17	0.12
S_2	1	0.8	0.7	0.6	0.5	0.45	0.4	0.4	0.3	0.2	0.1
S_3	1	0.9	0.8	0.7	0.5	0.4	0.3	0.2	0.2	0.1	0.1

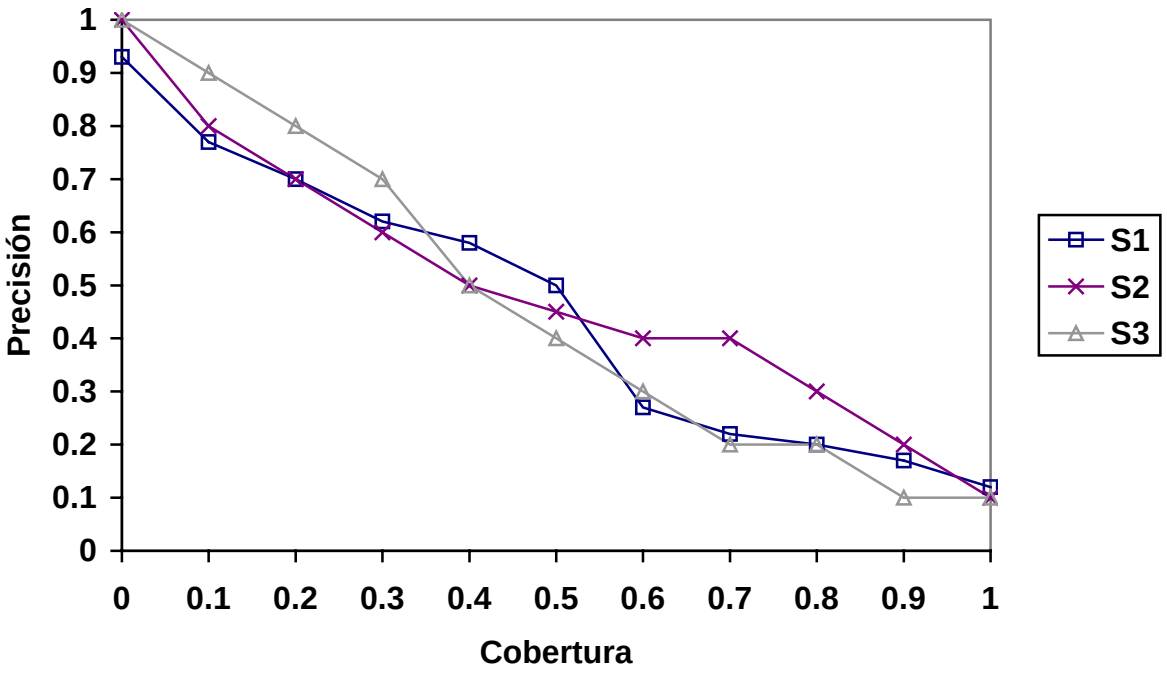


Figura 2-2. Curvas precisión/cobertura.

2.1.2 Procesamiento de Lenguaje Natural en IR

El Procesamiento de Lenguaje Natural (NLP, *Natural Language Processing*) ofrece técnicas y herramientas para abordar los problemas de ambigüedad léxica, variación terminológica y multilingüismo que afectan a la Recuperación de Información (IR).

El Procesamiento del Lenguaje Natural ha acompañado a la Recuperación de Información desde sus comienzos. Como conjunto de técnicas para procesar el contenido de los textos, el procesamiento lingüístico promete representaciones más ricas permitiendo indexaciones que mejoraran la recuperación. Es en la indexación de textos donde más se ha intentado aplicar el Procesamiento de Lenguaje Natural, en especial las técnicas de procesamiento sintáctico y semántico.

A continuación se enumeran brevemente los recursos y herramientas de utilidad para la tarea de IR, y en especial los utilizados a lo largo de este trabajo.

2.1.2.1 Redes léxico-semánticas: EuroWordNet

EuroWordNet (Vossen 1998) es una base de datos léxica multilingüe con relaciones semánticas entre las palabras de varios idiomas europeos: inglés, holandés, español, italiano, alemán, francés, checo y estonio. Esta base de datos se estructura en forma de redes semánticas formadas por unidades denominadas synsets. Un synset es un conjunto de sinónimos que corresponden a un mismo concepto. Entre los synsets se establecen las relaciones semánticas fundamentales: hiponimia/hiperonimia, meronimia, implicatura, etc.

El carácter multilingüe de EuroWordNet requiere una estructura adicional que permita interconectar los synsets de idiomas diferentes. Esta estructura es un Índice Interlingua (ILI) (Vossen 1999) que representa una lista no estructurada de conceptos (ILI-records) independiente del idioma. Estos conceptos constituyen un superconjunto de los conceptos que aparecen en las distintas redes de cada idioma. Por ejemplo, dos synsets pertenecientes a dos lenguas distintas, ligados mediante la relación de EQ_SYNONYM a un mismo ILI-record se consideran equivalentes en su significación semántica.

EWN es un recurso que se puede utilizar en tareas de Recuperación de Información como por ejemplo:

- identificar nueva terminología a partir de sinónimos,
- relacionar nueva terminología con terminología ya existente de cara a la recuperación de información,
- traducir consultas,
- expandir y refinar consultas mediante sinónimos, hipónimos e hiperónimos.

2.1.2.2 Diccionario monolingüe

Un diccionario monolingüe permite identificar palabras correctas e incorrectas (errores tipográficos y de ortografía) en los documentos de la colección. Además permite identificar expresiones equivalentes y enriquecer otros recursos léxicos como EuroWordNet (Gonzalo 1998a), (Peñas 2000), (Chugur 2001).

2.1.2.3 Diccionario bilingüe

Un diccionario bilingüe permite traducir las palabras de la consulta y ayudar a obtener correspondencias entre los términos de dos idiomas diferentes. Además ha servido para identificar enriquecer otros recursos léxicos como EuroWordNet.

2.1.2.4 Tokenizador

Un tokenizador sirve para separar oraciones, palabras y signos de puntuación de forma que sea posible su posterior tratamiento mediante herramientas como el analizador morfológico y el etiquetador de categorías gramaticales.

2.1.2.5 Analizador morfológico

Un analizador morfológico toma un texto tokenizado y asigna a cada palabra todos los lemas y categorías gramaticales que se ajustan a su morfología. El analizador morfológico utilizado en este trabajo (MACO) (Carmona 1998) ha sido desarrollado en la Universidad de Barcelona y la Universidad Politécnica de Cataluña. La *Figura 2-3* muestra la salida que produce el analizador morfológico tras el procesamiento del texto de entrada. Esta salida será la entrada del etiquetador o desambiguador de categoría gramatical.

2.1.2.6 Etiquetador de categorías gramaticales

Un etiquetador de categorías gramaticales es una herramienta que decide con qué categoría gramatical se está utilizando una unidad léxica en un determinado

contexto. La desambiguación de la categoría gramatical determina a su vez cuál es el lema correcto. El etiquetador utilizado en este trabajo (Relax) (Padró 1998) coordinada con el analizador morfológico. Al igual que MACO, también Relax ha sido desarrollado por la Universidad Politécnica de Cataluña. Relax realiza la desambiguación morfosintáctica sobre la salida de MACO, es decir, para cada una de las palabras se elige una sola etiqueta y su lema correspondiente entre todas las etiquetas y lemas que le proporciona el analizador morfológico como entrada.

Aunque la precisión de los etiquetadores es suficientemente elevada (superior al 80% de precisión), los etiquetadores son herramientas de elevado coste computacional lo que imposibilita su uso extensivo en la indexación de la web.

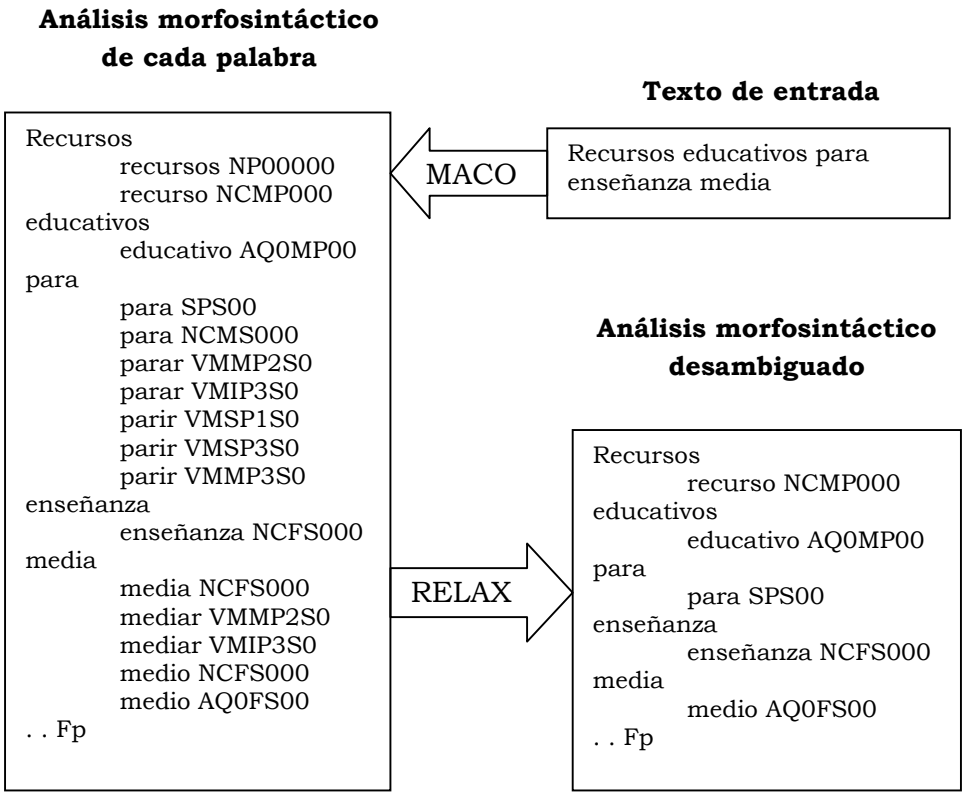


Figura 2-3. Encadenamiento del procesamiento morfosintáctico

2.1.2.7 Desambiguador del sentido de las palabras

Para elegir las palabras de expansión y traducción de una consulta o un texto, es conveniente conocer el sentido con el que se utilizan las palabras en un determinado contexto. Un desambiguador del sentido aborda este problema,

asignando a cada una de las palabras de un texto uno de sus sentidos contemplados en un diccionario.

Conocer el sentido de una palabra permite, además, situarlo en una red semántica y activar el concepto adecuado. Esto permite una indexación conceptual basada en los nodos de la red semántica.

Sin embargo, la desambiguación del sentido de las palabras, además de ser un proceso computacionalmente muy costoso, aún no ha alcanzado un buen nivel de resultados. En la conferencia Senseval-2 celebrada en el 2001 participaron 35 grupos y se presentaron 90 sistemas. El apartado más representativo correspondió a la de desambiguación de todas las palabras inglesas (*English All Words, Figura 2-4*) que incluye un 18% de palabras monosémicas. En este apartado, el mejor sistema alcanzó un 69% de palabras desambiguadas correctamente del total de palabras (*recall*). Este resultado no está muy alejado del 74% de coincidencia entre etiquetadores humanos, pero se alcanza únicamente con sistemas supervisados. La desambiguación mediante la selección del sentido más frecuente proporciona alrededor del 60% de *recall* lo cual resulta inalcanzable incluso para los mejores sistemas no supervisados (Fernández-Amorós 2002) (*UNED-AW-U2*, en la gráfica).

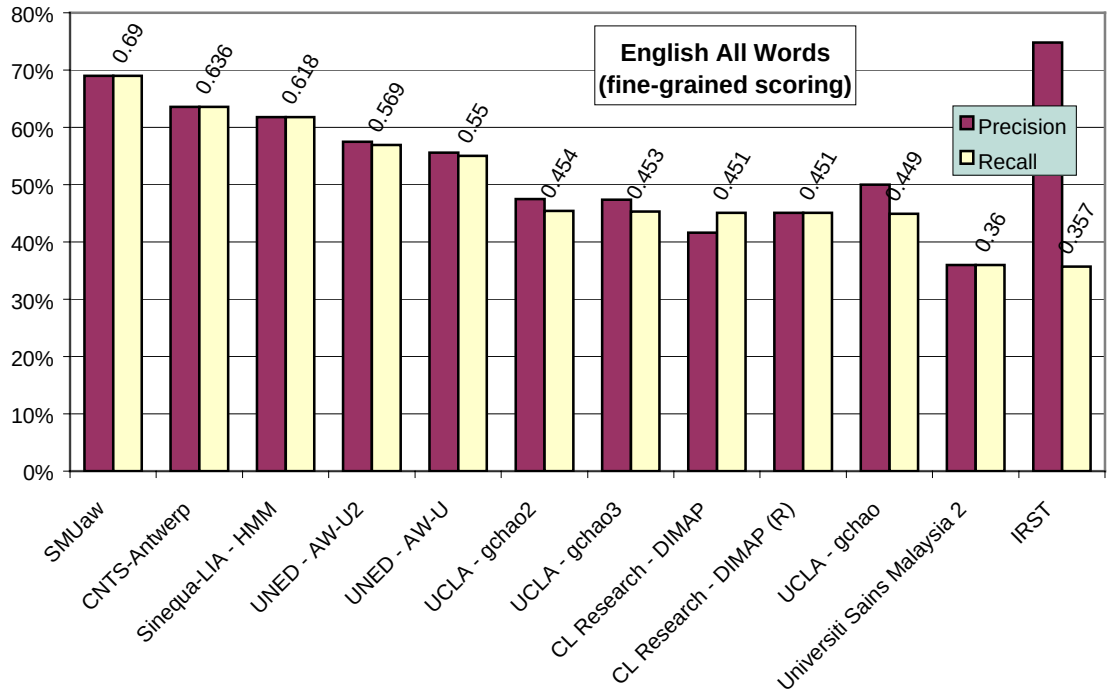


Figura 2-4. Resultados de WSD en Senseval-2

2.1.2.8 Analizador sintáctico superficial

El reconocimiento de unidades del discurso superiores a la palabra requiere la consideración de reglas sintácticas. El análisis sintáctico de una expresión permite su normalización y conceptualización, pero el mero reconocimiento de expresiones sintagmáticas no requiere un análisis tan complejo. Basta, por ejemplo, un análisis superficial del texto basado en el reconocimiento de secuencias adecuadas de categorías gramaticales.

2.1.3 Recuperación translingüe de información

Frecuentemente los usuarios son capaces de leer varios idiomas y, por tanto, es deseable que un usuario pueda realizar una consulta en un idioma de su elección y acceder a la información que proviene de documentos escritos en otro idioma diferente.

Por otra parte, usuarios que no comprendan más de un idioma también pueden beneficiarse de una búsqueda translingüe si se les proporciona una traducción de los documentos recuperados al idioma de la consulta. En algunas ocasiones basta la información de que el documento existe, en otras, es posible que la información que se busca no sea puramente textual pero que sin embargo se identifique textualmente, como en el caso de imágenes o archivos de sonido descritos mediante texto.

El problema de recuperar documentos en un idioma diferente a la consulta (recuperación translingüe) se aborda desde hace años en diversos foros de evaluación como el Text REtrieval Conference (TREC) o el Cross-Language Evaluation Forum (CLEF).

La búsqueda translingüe puede realizarse mediante vocabularios controlados o sobre texto libre:

- Búsqueda mediante vocabularios controlados. Fundamentalmente se utilizan los términos de un tesauro multilingüe como elementos de indexación de los documentos. Esto exige, una asignación que generalmente es manual, si bien puede llegar a ser automática tras entrenar un sistema de clasificación. La asignación de un término en un idioma implica la asignación de los términos apropiados en el resto de idiomas del tesauro lo que permite una búsqueda multilingüe.

- Búsqueda sobre texto libre. Los elementos de indexación proceden del propio texto. En general son palabras normalizadas (lemas o stems) aunque pueden ser sintagmas o conceptos obtenidos tras un procesamiento lingüístico de los textos.

En general, los foros de evaluación se centran en la recuperación translingüe sobre texto libre, siendo la aproximación más común la traducción de la consulta a los idiomas de la colección. Sin embargo, la opción más utilizada en Bibliotecas Digitales para tratar el problema del multilingüismo sigue siendo el uso de tesauros y vocabularios controlados.

Tratándose de texto libre, la recuperación translingüe acarrea problemas cuya resolución no es todavía satisfactoria. Estos problemas son fruto de la necesidad de procesar la consulta, los documentos o ambos, para que la información de uno pueda referirse al otro. Las formas generales de aproximar el idioma de la consulta y de los documentos son las siguientes:

1. *Traducir la consulta a los idiomas presentes en la colección de documentos*. La traducción de la consulta tiene como ventaja que al tratarse de un texto muy corto, el coste computacional de traducir la consulta a diversos idiomas es relativamente bajo. Sin embargo, esta característica de las consultas también supone su mayor inconveniente: la escasa información de contexto dificulta en gran medida una traducción de la consulta con niveles aceptables de precisión.

    - *Diccionarios*. Básicamente, se realiza la traducción de cada palabra de la consulta por las traducciones de un diccionario bilingüe. Esta aproximación reduce la efectividad entre un 40% y un 60% respecto a la búsqueda monolingüe (Ballesteros 1996), (Hull 1996). Sin embargo, es posible mejorar estos resultados considerando sintagmas (Ballesteros 1998), agrupando sinónimos mediante operadores de consulta (Pirkola 1998), o considerando únicamente las traducciones que pueden volver a traducirse (traducción inversa) por el término de partida (Boughanem 2002).

    - *Traducción automática*. La traducción se realiza mediante alguna aplicación de traducción automática. Los resultados dependen del par de idiomas implicados pero, en general, son peores cuanto más breve es la consulta (caso de los buscadores en la web).

    - *Corpus paralelo*. Partiendo de un corpus en el que cada documento tiene asociado otro documento traducción del primero, es posible

aplicar técnicas automáticas para traducir la consulta. Por ejemplo, es posible extender al caso translingüe los modelos de pseudo-relevance feedback, espacio vectorial generalizado (GVSM) o Latent Semantic Indexing obteniendo resultados muy próximos al caso monolingüe (Carbonell 1997).

- *Corpus comparable*. La dificultad de disponer de un corpus paralelo lleva a explorar aproximaciones basadas en corpus con documentos similares (no traducciones) en diferentes idiomas. La dificultad se traslada a la creación automática de un corpus comparable pero, en determinados dominios, los resultados en búsqueda translingüe pueden ser cercanos al caso monolingüe.

2. *Traducir la colección de documentos a los posibles idiomas de consulta* (Oard 1998). Debido a la extensión de los documentos, su traducción al idioma de la consulta puede realizarse con mayor precisión. Además, el proceso de traducción sólo se realiza una vez aunque deba realizarse para cada una de las lenguas posibles de consulta. Sin embargo, cuando se trabajan con grandes colecciones multilingües en un entorno en el que las consultas pueden realizarse en varios idiomas, el coste computacional puede ser muy elevado. Este es el caso del acceso a la información en el ámbito de las lenguas europeas y de la información contenida en Internet.

3. *Traducir consulta y documentos a un mismo idioma*. La traducción tanto de consultas como documentos a un tercer idioma tiene la ventaja de necesitar únicamente recursos léxicos bilingües respecto a una sola lengua. En general, resulta difícil disponer de diccionarios bilingües para cada par de lenguas que puedan encontrarse en el sistema de recuperación. Por otra parte, los documentos únicamente deben traducirse una vez a un único idioma, reduciéndose el coste computacional. El inconveniente es que se multiplica el ruido que producen las traducciones, llegando a una grave distorsión de la consulta.

4. Traducir recíprocamente consulta y documentos, considerando ambas traducciones simultáneamente (McCarley 1999). Los resultados se pueden mejorar si se realizan y consideran ambas traducciones simultáneamente, la de la consulta al idioma de los documentos y la de los documentos al idioma de la consulta.

5. *Traducir ambos a una misma representación conceptual* (Gonzalo 1998b). Una última posibilidad es traducir tanto consultas como documentos a un lenguaje intermedio o representación conceptual utilizando ontologías o

redes semánticas multilingües. En algunas redes semánticas, las palabras de los diferentes idiomas están ligadas a una misma representación, lenguaje intermedio o Interlingua. De esta forma, se dispone de un vehículo para ir de un idioma a otro a través de esta representación conceptual. Sin embargo, debido a que los nodos de estas redes son conceptos, es necesario determinar con qué sentido se utilizan las palabras para saber a qué nodos corresponden. Esto añade ventajas e inconvenientes al proceso de recuperación de información. La principal ventaja es que permite la recuperación de documentos que contengan palabras sinónimas o con alguna otra relación semántica. El mayor inconveniente al uso de redes semánticas es que se hace necesaria la desambiguación previa del sentido de las palabras (WSD, *Word Sense Disambiguation)*. Esto supone un problema todavía abierto y añade un elevado coste computacional al sistema. El uso de redes semánticas introduce otra dificultad en la traducción, y es que las unidades de traducción son más amplias que las unidades léxicas que se consideran en las redes semánticas y esta pérdida de información en la representación conceptual afecta negativamente a la recuperación (Verdejo 2000).

2.2 Indexación de sintagmas en IR

Ya desde los primeros trabajos de Recuperación de Información, los vocabularios controlados se han utilizado para indexar textos (Lancaster 1972;Hutchins 1975). Estos vocabularios, como es natural, contienen tanto términos simples como compuestos (sintagmas). Sin embargo, la rigidez de los términos en estos vocabularios controlados impone limitaciones en su uso como elementos de indexación.

Las técnicas aplicadas a la identificación y uso de sintagmas y grupos nominales se ha estudiado profusamente en las últimas décadas por dos razones: por una parte, los sintagmas son relativamente fáciles de obtener y, por otra, no hay una única definición concreta y específica de lo que es un sintagma. Históricamente, un sintagma en recuperación de información ha sido cualquier cosa desde n-gramas validados estadísticamente hasta estructuras sintácticas como secuencias de palabras con determinadas categorías gramaticales.

A pesar del gran número de experimentos que han intentado corroborar si el uso de sintagmas permite mejorar la efectividad de la recuperación, los resultados no son claros y a veces, incluso, son contradictorios (Pickens 2000;Gonzalo 1999a). Del trabajo de (Pickens 2000), sin embargo, se puede concluir que la adyacencia de las

palabras de la consulta en un documento supone una evidencia positiva en favor de la relevancia del documento y, por tanto, que una consideración adecuada de sintagmas debería llevar a un aumento en la precisión de la recuperación.

En este sentido, las palabras que aparecen en mayor número de compuestos diferentes (términos con mayor dispersión léxica) tienden a ser conceptos clave de sus respectivos dominios (Anick 1999). Por ser palabras muy frecuentes en la colección tienen escaso valor discriminatorio por ellas mismas. Sin embargo, no ocurre así si se consideran los sintagmas en los que interviene.

Desde el comienzo, el procesamiento automático se ha dirigido a conseguir el mismo tipo de indexación que se conseguía de forma manual (Salton 1968;Bely 1970). El trabajo de Bely tenía como objetivo la identificación automática de expresiones que tuvieran relación con los conceptos de un tesauro. Si bien no llevaron a cabo ningún tipo de evaluación respecto a la mejora en la recuperación de información, sí encontraron que las descripciones automáticas eran bastante similares a las manuales. El trabajo de Salton también analiza los términos de un tesauro dándoles una representación normalizada (sintagmas sintácticos) y comparando su empleo en la recuperación respecto al empleo de sintagmas estadísticos. Los sintagmas estadísticos fueron definidos en esta ocasión como co-ocurrencia, dentro de los límites de una oración, de los constituyentes de un término del tesauro. La conclusión de Salton fue que el empleo de sintagmas sintácticos no proporcionaban mejores resultados en la recuperación que el uso de sintagmas estadísticos.

También (Cleverdon 1967) y (Salton 1972) mostraron que se obtenían resultados similares utilizando lenguajes controlados con términos complejos que utilizando descripciones en lenguaje natural ya fueran de términos complejos como de términos simples.

(Dillon 1983) sin embargo, concluyó que el empleo de sintagmas seleccionados y normalizados tras un procesamiento lingüístico proporcionan ligeras mejoras en la recuperación respecto a la utilización de términos simples.

(Fagan 1989) evaluó la utilización de sintagmas no sintácticos atendiendo a parámetros como el grado de proximidad así como la frecuencia de aparición de los constituyentes. La noción de proximidad relaja considerablemente el concepto de sintagma, resultando en una mejora de la recuperación respecto a la utilización tanto de sintagmas sintácticos como de términos simples. Sus conclusiones fueron que no merecía la pena considerar compuestos con más de dos constituyentes y que aquellos compuestos con más de dos constituyentes debían descomponerse en compuestos de dos. Asimismo, los compuestos no debían sustituir a las componentes sino que debían considerarse tanto los compuestos como los

elementos de los compuestos. Según Fagan, una de las razones para que los términos compuestos no puedan contribuir a mejorar la recuperación es simplemente que no aparecen en las consultas y, por tanto, su presencia en los documentos se ignora.

También (Croft 1991) llega a la conclusión de que deben pesarse tanto el compuesto como los elementos constituyentes del compuesto.

Estos primeros experimentos muestran que las técnicas lingüísticas aplicadas a la indexación de textos no permiten mejorar la recuperación de información. Sin embargo, puede argumentarse que son experimentos con poca relevancia estadística dado el pequeño volumen de las colecciones de evaluación. Además, los textos indexados son *abstracts* en los que el grado de significación y precisión de todos sus términos es alto y en cambio es baja su frecuencia de aparición. Es decir, su comportamiento no se ajusta con el de los textos cuando se consideran documentos completos.

En la última década los avances técnicos hicieron posible la consideración de los textos completos y esto, junto con la mejora de las técnicas NLP, llevó a reconsiderar el papel que desempeñan las técnicas lingüísticas en la recuperación de información.

(Krovetz 1997) volvió a estudiar el papel que puede desempeñar la indexación de sintagmas en la recuperación. Al igual que en experimentos anteriores, sus resultados muestran que la detección de colocaciones puede ser ligeramente beneficiosa para la recuperación, si bien resulta indispensable considerar las componentes de la colocación asignándoles también un peso parcial.

(Sparck Jones 1999) traslada a las experiencias TREC (Text Retrieval Conference) de los últimos años, la revisión de conceptos y presupuestos al respecto de la utilidad de las técnicas lingüísticas aplicadas a la indexación de textos, para concluir que no producen mejoras significativas en la recuperación de información.

Es en el uso de vocabularios controlados donde mayor aplicación tienen las técnicas lingüísticas. No sólo en la indexación, sino también en otras tareas como la categorización de textos. Los trabajos de Jacquemin (Jacquemin 2000) muestran el uso de técnicas lingüísticas (análisis morfológico, etiquetado de categoría gramatical y análisis sintáctico superficial) para abordar los problemas de variación terminológica tanto morfosintáctica como semántica. De esta forma Jacquemin consigue mejorar los índices de recuperación (recall) de textos indexados con vocabularios controlados.

2.3 Ambigüedad léxica en IR

Para traducir de forma precisa una palabra, es necesario determinar su categoría gramatical, lema y sentido en el diccionario. Las técnicas automáticas de desambiguación léxica permiten abordar estas tareas y considerar términos de indexación desambiguados. Los siguientes trabajos han aplicado estas técnicas de desambiguación léxica a los índices de recuperación sin resultados concluyentes.

(Voorhees 1993) ideó un método de desambiguación basado en las relaciones semánticas de la jerarquía de WordNet. Aplicado a la indexación de colecciones como CACM (Salton 1983b) tuvo como resultado un empeoramiento en la recuperación. Su evaluación cualitativa le llevó a la conclusión de que los malos resultados en la recuperación se debían a la falta de precisión en la desambiguación del sentido de las palabras.

También (Sussna 1993) utilizó un desambiguador basado en las relaciones semánticas de WordNet obteniendo malos resultados en la recuperación. En este caso Sussna sí evaluó cuantitativamente la precisión del desambiguador. Para ello seleccionó y desambiguó manualmente 320 ocurrencias de palabras ambiguas. Jugando con el tamaño de la ventana de contexto obtuvo un máximo del 56% de precisión en la desambiguación. Con estos resultados no es posible determinar si el empeoramiento en la recuperación se debe a la falta de precisión en la desambiguación o a que la desambiguación no es una buena estrategia para la recuperación.

(Wallis 1993) utilizó un desambiguador basado en las definiciones del diccionario LDOCE y tampoco obtuvo buenos resultados sobre colecciones como CACM.

(Richardson 1995) utilizaron un sistema de recuperación basado en similitud conceptual que también resultó en un empeoramiento de la recuperación. Los autores no pudieron determinar en qué grado el problema se debía a una estrategia ineficaz o a los errores introducidos en la desambiguación del sentido de las palabras (WSD, *Word Sense Disambiguation*).

Sin embargo, (Smeaton 1996) con una estrategia parecida pero con desambiguación manual y sobre textos muy breves (pies de imagen), sí obtuvieron una mejora en la recuperación.

(Voorhees 1994) realiza una expansión manual sobre las consultas del TREC, con palabras semánticamente relacionadas mediante WordNet, obteniendo sólo ligeras mejoras en el caso de las consultas más cortas.

Los experimentos de (Krovetz 1997) consideraron la categoría gramatical de las palabras en la indexación determinando una pérdida de efectividad en la recuperación. Sin embargo, Krovetz no pudo determinar en qué grado la pérdida de efectividad se debía al efecto de considerar la categoría gramatical de las palabras, y en qué grado los resultados se debían a los errores de desambiguación en el proceso automático. Respecto a la discriminación de los sentidos de las palabras considerando su categoría gramatical anotada con el etiquetador de Church, Krovetz también obtuvo un empeoramiento en la recuperación. Krovetz señaló que más de la mitad de las palabras de un diccionario con diferente categoría gramatical tenían, sin embargo, significados relacionados. A pesar de esto, Krovetz tampoco pudo decidir si el empeoramiento de la recuperación se debía a esta pérdida de relaciones semánticas o se debía a los errores del etiquetado automático.

Ya con anterioridad, y sobre un sistema de tradicional de recuperación, (Krovetz 1992) estudiaron si las palabras de la consulta se utilizaban o no con el mismo sentido tanto en la consulta como en los documentos recuperados en lo alto del ranking. Comprobaron que, efectivamente, en la mayoría de los casos el sentido era el mismo. Para explicar este fenómeno dieron dos razones. La primera es que el conjunto de palabras de la consulta impone implícitamente restricciones sobre el sentido de cada una de ellas. Por ejemplo, una consulta con las palabras "*banco moneda fisco"* determina que los primeros documentos recuperados (que contendrán *moneda y fisco*) tengan el mismo sentido de *banco* que la consulta. La segunda razón es que, en la mayoría de las consultas (75.6% en la colección CACM), las palabras son monosémicas o se utilizan con un sentido muy frecuente (sentido en más del 80% de las ocurrencias de la palabra) y, por tanto, los errores resultan proporcionalmente escasos. Sin embargo, según los autores se puede producir hasta un 33% de mejora en la precisión media en colecciones como CACM si se realizara una correcta desambiguación de palabras con sentidos uniformemente distribuidos o utilizadas con sentidos minoritarios. Este aumento de precisión resulta de interés en búsquedas con un alto *recall*.

El problema de discernir el efecto que produce en la recuperación la consideración de sentidos, del efecto que produce la falta de precisión en el proceso de desambiguación automática fue abordado por (Sanderson 1994). Para ello, Sanderson creó pseudo-palabras de tamaño n sustituyendo n palabras por la cadena que resulta de concatenarlas. Por ejemplo, tanto *banana* como *kalashnikov* se sustituyen por la pseudo-palabra de tamaño 2 *banana/kalashnikov*. De esta forma, la pseudo-palabra se puede desambiguar con un 100% de precisión sustituyéndola por la palabra original (en nuestro ejemplo *banana* o *kalashnikov*). Sanderson obtuvo unos resultados que mostraban que el proceso de Recuperación de Información se comporta de forma similar con independencia de la longitud de las pseudo-palabras, es decir, según Sanderson es un proceso resistente al

aumento en el grado de ambigüedad. Por otro lado, Sanderson comprobó la sensibilidad de la recuperación en la introducción de errores, llegando a la conclusión de que si se desambiguaba el sentido de las palabras con una precisión inferior al 90% la recuperación era peor que si no se desambiguaba en absoluto. Todo esto llevaba a la conclusión de que era preferible no realizar desambiguación alguna salvo si se hacía con mucha precisión.

La cuestión que quedó abierta fue si la ambigüedad de las palabras reales es comparable a la ambigüedad de las pseudo-palabras. De ser así, la desambiguación sin más del sentido de las palabras no aportaría mejoras significativas a la recuperación.

(Gonzalo 1998b) construyeron una colección de prueba en la que tanto consultas como textos fueron anotados manualmente con etiquetas de categoría gramatical y etiquetas semánticas correspondientes al sentido de las palabras en WordNet. El etiquetado manual reduce al mínimo los errores de desambiguación y crea unas condiciones óptimas para dilucidar si una indexación semántica puede mejorar o no la recuperación de información. El objetivo de esta aproximación era traducir las palabras a una representación conceptual idéntica para todas las palabras que se utilizan con el mismo significado. Para ello, todas las palabras utilizadas con el mismo significado se traducen a un mismo índice conceptual denominado *synset*. Un synset es el índice de un conjunto de palabras que son sinónimas para el significado asociado al synset.

En este marco, los experimentos de (Gonzalo 1998b) ofrecen unos resultados que muestran una mejora significativa de la recuperación incluso para consultas largas. También muestran que la indexación mediante synsets de WordNet proporciona mejores resultados de recuperación que la indexación mediante sentidos.

A diferencia de la mera desambiguación del sentido de las palabras que sólo permite recuperar documentos con la misma palabra y mismo sentido, la indexación mediante synsets permite recuperar documentos con palabras diferentes siempre que se utilicen con el mismo significado. Además, mantener el identificador de synset como índice de recuperación en lugar de realizar una expansión con las palabras que componen el synset elimina el ruido que introducen los significados adicionales que tienen las palabras. Estas particularidades asociadas a la indexación conceptual basada en synsets de WordNet podrían explicar la mejora en la recuperación, a diferencia de los estudios de (Voorhees 1994) sobre expansión y de (Sanderson 1994) sobre desambiguación.

En (Schütze 1995) sí se consiguió una mejora en la recuperación textual tras la discriminación de sentidos extraídos de una colección de test sobre la consideración

de matrices de co-ocurrencia (tesauros automáticos). Sin embargo, cuando se utilizan los sentidos fijados en un diccionario, tesauro u ontología (e.g. WordNet) la discriminación de sentidos no parece mejorar significativamente la recuperación si no se hace con un nivel alto de precisión.

2.4 Exploración de términos en el acceso a la información

La consideración de sintagmas no sólo abre la posibilidad de añadir términos a los índices de recuperación, sino que proporciona otra posibilidad en el acceso a la información: en lugar de navegar únicamente por los documentos de la colección es posible navegar por la terminología de la colección y, a partir de ella, acceder a los documentos que puedan ser relevantes para el usuario. Existen varias aproximaciones que intentan explotar las posibilidades de interacción mediante la sugerencia o exploración de términos (palabras y sintagmas):

1. De construcción manual
   - Jerarquías temáticas como la de (Yahoo).
   - Tesauros como el utilizado en (ERIC), etc.
2. De construcción automática
   - Jerarquías extraídas de forma automática mediante la agrupación automática de documentos (clustering) en clases anidadas (Hearst 1996) o mediante relaciones de subsunción entre términos (Sanderson 1999).
   - Jerarquías de sub-sintagmas (Nevill-Manning 1999), (Paynter 2001b).
   - Enlaces entre documentos con palabras claves similares extraídas automáticamente (Jones 1999), (Frank 1999).
   - Expansión de la consulta mediante sintagmas sugeridos por el sistema (Anick 1999).

Antes de discutir todas estas aproximaciones es necesario destacar que ninguna de ellas aborda los problemas de variación morfosintáctica, semántica y translingüe de los términos.

2.4.1 Jerarquías temáticas

En este tipo de sistemas el usuario navega por una jerarquía de materias hasta el grupo de documentos que son de su interés. El usuario está obligado a identificar la información que le interesa a través de los términos de clasificación. Este tipo de sistemas requiere la clasificación previa de los documentos de acuerdo con el vocabulario, tarea costosa que no siempre resulta sencilla y que en la mayoría de

los casos se realiza de forma manual. A modo de ejemplo, la *Figura 2-5* muestra el nivel más alto de la jerarquía temática de Yahoo en español.

Arte y cultura Literatura, Teatro, Museos...	**Internet y ordenadores** WWW, Aplicaciones, Chat...
Ciencia y tecnología Animales, Informática, Ingeniería...	**Materiales de consulta** Bibliotecas, Diccionarios...
Ciencias sociales Economía, Psicología, Historia...	**Medios de comunicación** Temas de actualidad, Periódicos, TV...
Deportes y ocio Fútbol, Deportes, Turismo...	**Política y gobierno** Países, Embajadas, Derecho...
Economía y negocios Para empresas, Para consumidores, Empleo...	**Salud** Medicina, Enfermedades...
Educación y formación Primaria, Secundaria, Universidades...	**Sociedad** Gastronomía, Culturas, Religión...
Espectáculos y diversión Cine, Actores, Música, ¡Genial!...	**Zonas geográficas** Países, Europa, España, CC.AA....

Figura 2-5. Nivel superior de la jerarquía temática de Yahoo en español.

2.4.2 Exploración mediante listas y tesauros

Cuando los documentos han sido clasificados de acuerdo con un vocabulario controlado, es posible utilizar ese vocabulario como vía de acceso a la colección.

Una de las maneras de navegar por un vocabulario controlado es mediante listas estructuradas de términos predefinidos que cubren todas las ideas y materias que aparecen en textos de un determinado dominio. Los documentos se organizan bajo dichos términos y se puede navegar por la jerarquía de términos a través de relaciones de especificidad (NT, BT y RT).

También es usual que se permita una búsqueda en listas de indexación ordenadas alfabéticamente de los términos contenidos en campos como autor, título o materia con los que se ha clasificado el documento. Primero se le pide al usuario que introduzca un término y a continuación se le muestra una lista de los términos semejantes a él que están contemplados en la lista de indexación. El usuario entonces elige el término entre los predeterminados para obtener los documentos relacionados.

The Educational Resources Information Center (ERIC) es una buena referencia de las posibilidades que ofrece un tesauro en la exploración de colecciones de documentos. ERIC es un sistema de información diseñado para proporcionar acceso a una base documental de textos relacionados con la educación. Su construcción comenzó en 1966 a cargo del "Office of Educational Research and Improvement" del Departamento de Educación de Estados Unidos, y la Biblioteca Nacional de Educación de Estados Unidos. La base de datos contiene más de 950.000 resúmenes de documentos y artículos de revistas sobre investigación y prácticas educativas. La información de esta base de datos se actualiza mensualmente.

Entre los sistemas de acceso a la base de datos ERIC se pueden distinguir dos filosofías principales de búsqueda y recuperación de recursos:

1. *Free form searching*: búsqueda a partir de las palabras de consulta proporcionados por el usuario sin restricción alguna.

2. *Search Wizard*: búsqueda a partir de los términos de un tesauro en el que el sistema permite navegar por el tesauro para seleccionar los términos que van conformando interactivamente la consulta.

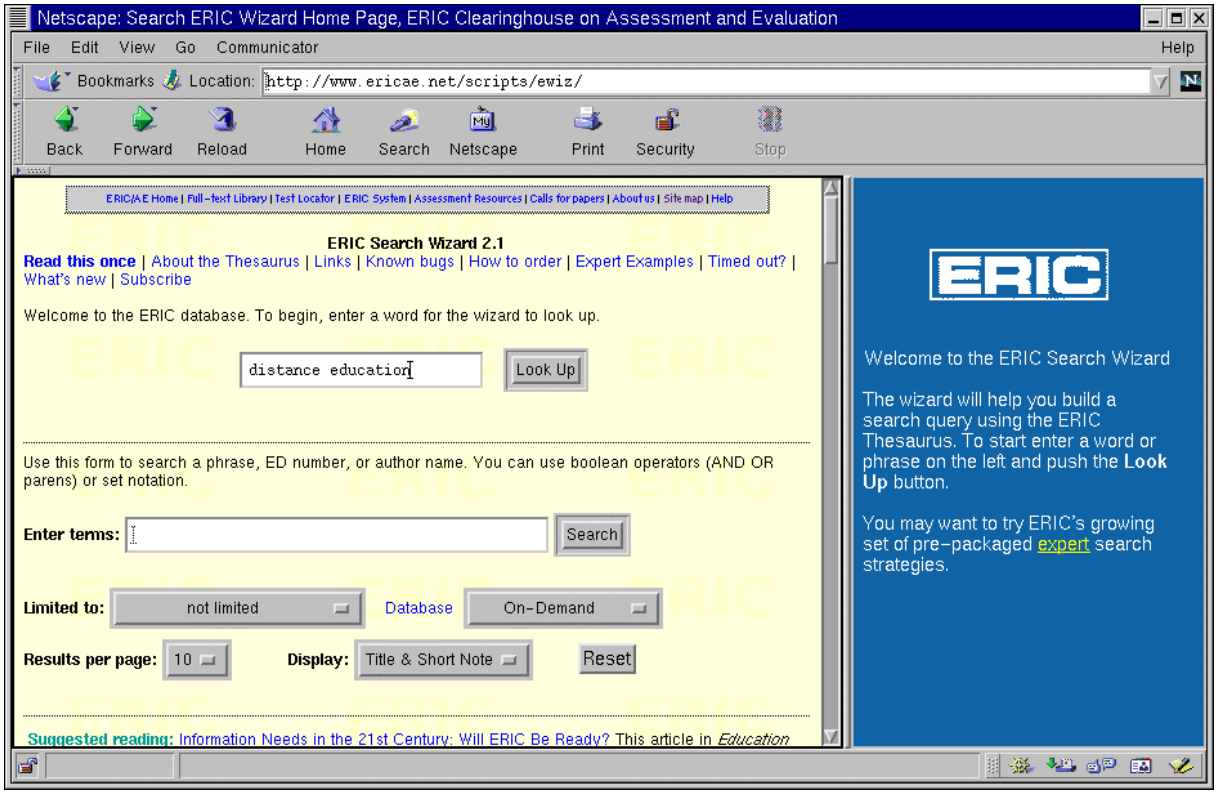


Figura 2-6. Introducción de términos iniciales en ERIC Wizard

La *Figura 2-6* y la *Figura 2-7* muestran el tipo de búsqueda en el que el usuario construye la consulta a partir de los términos del tesauro. En primer lugar el sistema ofrece al usuario la posibilidad de introducir un término libremente. El sistema busca en el tesauro la presencia de ese término. Si no está, el sistema lo notifica pero, siempre que sea posible, proporciona términos relacionados que sí pertenecen al tesauro y que permitirán al usuario navegar por él a partir de ese momento.

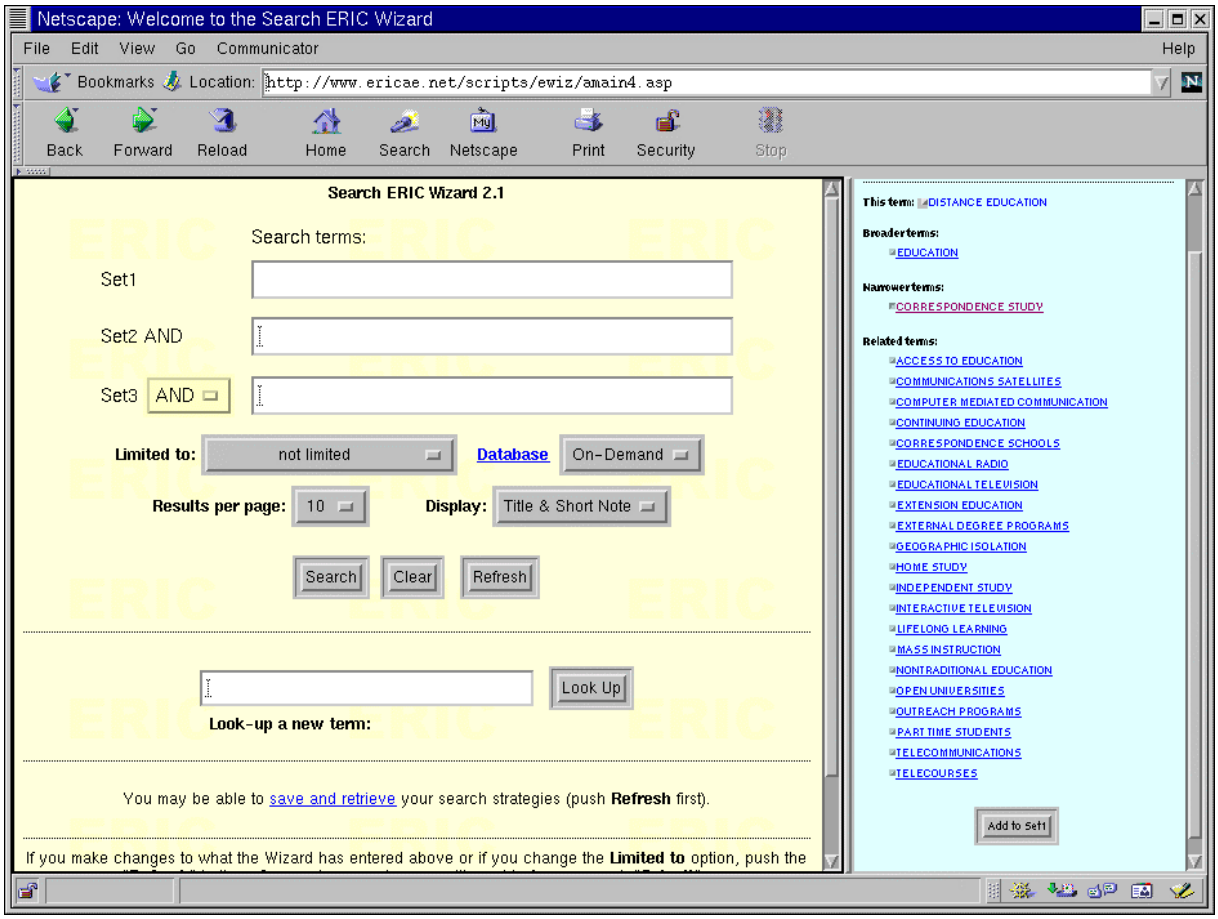


Figura 2-7. Selección de Términos de Consulta en ERIC Wizard a través del tesauro

El usuario puede añadir su término a la consulta aunque no pertenezca al tesauro, o bien puede seguir navegando por el tesauro hasta encontrar los términos precisos que incluir a la consulta. Por ejemplo, suponiendo que ha escrito o seleccionado "*distance education*", el usuario puede:

- Añadir este término a la consulta.
- Seleccionar términos más generales (*Broader Terms, BT*) como *"education"*, más específicos (*Narrower Terms, NT*) como *"correspondence study"* o

simplemente relacionados (*Related Terms, RT*) como “access to education”, “educational radio”, etc.

- Explorar a su vez los términos relacionados con estos otros términos.

De esta forma el usuario va construyendo la consulta. Una vez terminado este proceso se realiza la recuperación y se devuelve la lista de los documentos relevantes para la misma.

2.4.3 Agrupación automática de documentos en clases anidadas

La agrupación automática de documentos (clustering) en clases anidadas de (Hearst 1996) proporciona como descriptores de una agrupación el conjunto de palabras más características de la misma. El usuario, entonces, puede navegar por la jerarquía de clases en las que la secuencia de palabras asociada a cada clase da una idea del contenido de la agrupación. Sin embargo, esta secuencia de palabras obtenida artificialmente no suele corresponderse con un concepto bien definido para el usuario por lo que su interpretación no siempre resulta sencilla. Su falta de estructura lingüística dificulta su comprensión como concepto, así como su extensión a un dominio multilingüe.

2.4.4 Jerarquías de subsunción

Otra manera de navegar por términos es construir una jerarquía de los mismos de más generales a más específicos según una relación de subsunción. Sobre las ideas de (Forsyth R. 1986), (Sanderson 1999) define la relación de subsunción de la siguiente manera[6]: “el término *x* subsume al término *y* si el conjunto de documentos que contienen a *y* es un subconjunto de los documentos que contienen a *x*”. De esta manera, la *frecuencia de documentos* (*document frequency,* número de documentos que contienen al término) proporciona un mecanismo de ordenación de los términos de más generales a más específicos, permitiendo la construcción de jerarquías de términos. Sin embargo, el significado de este tipo de relación jerárquica es muy difícil de determinar.

El objetivo de Sanderson es construir jerarquías conceptuales pero, una vez más, uno de los problemas de esta aproximación es la ambigüedad léxica: ¿dónde situar un término polisémico en la jerarquía? La relación de subsunción no debería

[6] (Esta definición se puede expresar en términos de probabilidades como: P(x|y)=1, P(y|x)<1). Los autores relajan esta condición para obtener más relaciones: P(x|y)>=0.8, P(y|x)=1.

aplicarse sobre palabras sino sobre conceptos. Sin embargo, debido a que todavía no se dispone de técnicas suficientemente precisas para desambiguar semánticamente las palabras (WSD), Sanderson construye la jerarquía a partir de un subconjunto de documentos donde se espera que los términos se utilicen con el mismo significado: los primeros 500 documentos obtenidos como resultado de una consulta.

2.4.5 Expansión de la consulta mediante sintagmas

(Anick 1999) explota la tendencia de las palabras que suponen conceptos clave del dominio a participar en familias de compuestos léxicos semánticamente relacionados. La hipótesis de dispersión léxica enuncia que los conceptos clave dentro de una colección tienen mayor tendencia a participar en una amplia variedad de compuestos léxicos semánticamente relacionados. La dispersión léxica de una palabra se define como el número de compuestos diferentes en los que aparece dicha palabra, dentro de un determinado conjunto de documentos. El sistema *Paraphrase Search Assistant* sugiere al usuario los términos con mayor dispersión léxica en el conjunto de documentos recuperados, como términos para refinar la consulta.

La medida de dispersión léxica se ve afectada por documentos aislados con una gran cantidad de términos que no pertenecen al dominio. Por esta razón, los autores enriquecen la medida considerando la *frecuencia de documentos* (*document frequency*, número de documentos que contiene el término) con el fin ajustar los resultados. Sin embargo, el mayor problema de la dispersión léxica es que muchos conceptos clave son a su vez sintagmas lexicalizados que pierden su sentido al considerar sus componentes aisladas. El concepto de dispersión léxica no permite identificar estos conceptos clave expresados con sintagmas. De esta forma, las palabras identificadas mediante dispersión léxica son demasiado generales, aportan poca información al usuario y tienen poca capacidad de discriminación conceptual.

2.4.6 Navegación por sintagmas clave

Phrasier (Jones 1999) es un sistema que permite navegar entre documentos a través de *sintagmas clave*. Los autores utilizan una herramienta (*KEA*)(Frank 1999) de extracción de *sintagmas clave* para asignar 10 sintagmas a cada documento. Estos sintagmas identificados automáticamente se utilizan como términos de indexación del documento. De esta forma, a partir de los *sintagmas clave* es posible acceder a los documentos que lo contienen pero, además, los documentos que comparten

sintagmas clave quedan enlazados entre sí resultando posible navegar por la colección a través de estos enlaces. Además, se construye un espacio vectorial de *sintagmas clave* en el que las medidas de similitud entre vectores permiten ordenar una lista de los documentos relacionados con otro documento. Este sistema sólo permite navegar y explotar los sintagmas previamente extraídos mediante KEA, no siendo posible relacionar expresiones conceptualmente equivalentes.

2.4.7 Jerarquías de sub-sintagmas

El usuario de un sistema de exploración de sub-sintagmas (*phrase browsing*) introduce una palabra y el sistema le devuelve todos los sintagmas que contienen esa palabra. El usuario a su vez elige uno de estos sintagmas y el sistema le ofrece sintagmas más largos que lo contienen, o bien, la posibilidad de acceder a los documentos.

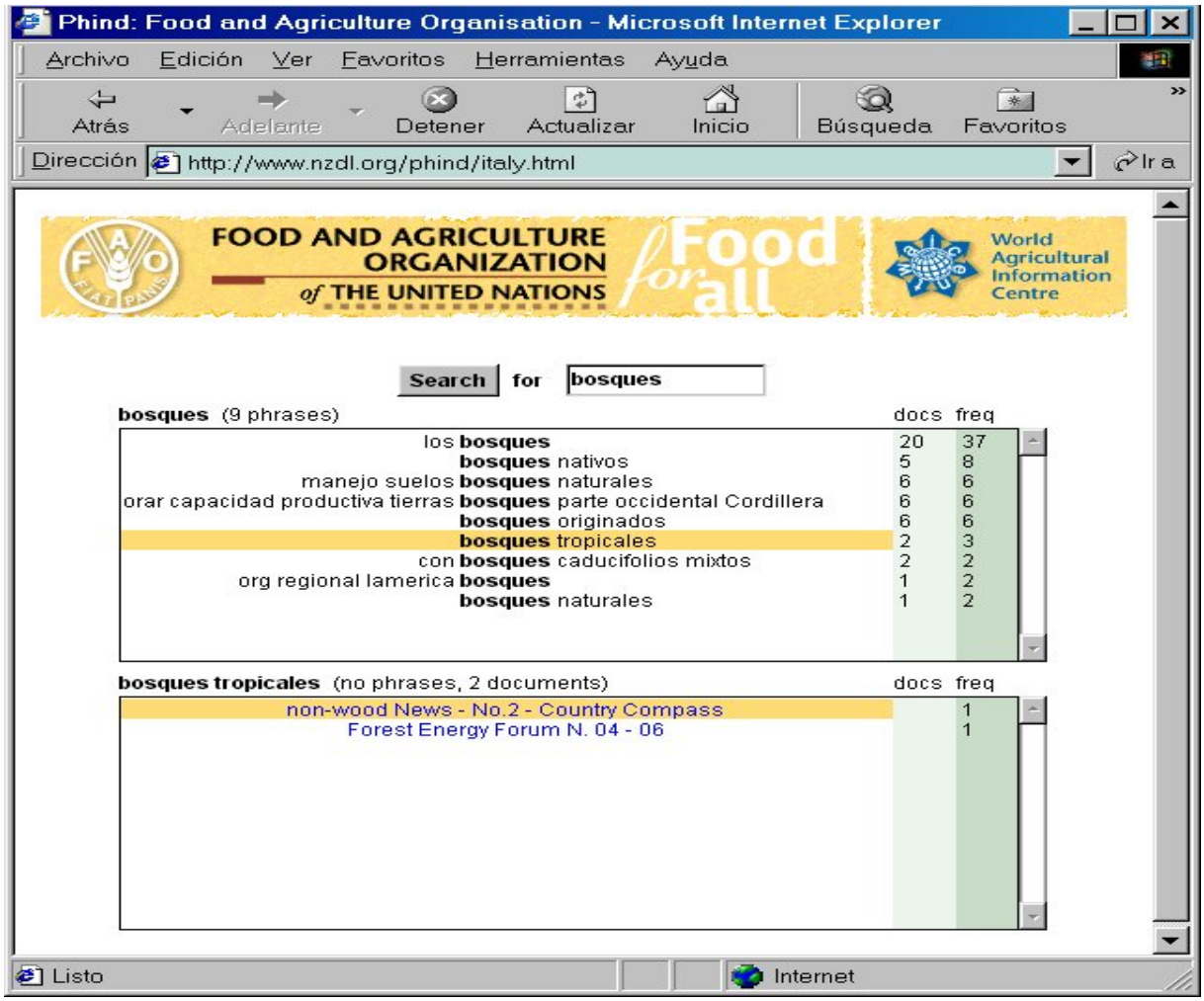


Figura 2-8. Exploración de sintagmas con el sistema Phind.

Uno de los sistemas más representativos de exploración basada en sub-sintagmas es *Phind* de la Universidad de Waikato (Nueva Zelanda) que se utiliza en el paquete de software para bibliotecas digitales *Greenstone* (Witten 1999a) desarrollado por dicha universidad. La *Figura 2-8* muestra un ejemplo de búsqueda mediante exploración de sintagmas en el sistema *Phind*.

En estos sistemas, la precisión con la que se han obtenido los sintagmas es importante, pues al usuario se le podría llegar a ofrecer demasiado información errónea o irrelevante que dificulte la identificación de la información que está buscando. Como se observa en el ejemplo, los sintagmas se definen como secuencias de palabras que ocurren más de una vez en la colección. Esto provoca que muchos de los sintagmas propuestos sean expresiones incorrectas. Además, el número de sintagmas que pueden contener una palabra puede llegar a ser realmente extenso, resultaría conveniente imponer algún tipo de restricción como por ejemplo la ocurrencia de una segunda palabra.

El ejemplo muestra también que debido a que no se utiliza conocimiento lingüístico, el sistema no puede tratar variación morfosintáctica, semántica ni translingüe. Algo relativamente sencillo como la exploración de los sintagmas que contienen la palabra *bosque* requiere una sesión diferente a la exploración de los sintagmas que contienen el plural *bosques*.

2.5 Conclusiones

2.5.1 Indexación con técnicas lingüísticas

A pesar de las expectativas que crean las técnicas lingüísticas a la hora de abordar problemas como la ambigüedad léxica y la variación terminológica, su aplicación a la indexación no redunda en un beneficio claro en la recuperación. Son muchos los factores que influyen en estos resultados y el número de variables que desempeña algún papel en la recuperación crece al considerarse un procesamiento lingüístico automático.

En un principio, los sintagmas obtenidos y normalizados mediante técnicas lingüísticas deberían comportarse mejor que los sintagmas obtenidos estadísticamente porque pueden tratar mejor problemas como los de variación terminológica. Sin embargo, los sintagmas extraídos y normalizados

lingüísticamente no dejan de ser representaciones bastante pobres del contenido semántico de los textos y puede ser esta la causa de que en recuperación de información no haya diferencias significativas en cuanto a su uso. Sólo se consigue cierta mejora en el uso de sintagmas cuando se relaja la restricción de adyacencia de los constituyentes, se considera su proximidad, y no se sustituyen las componentes por el compuesto sino que se consideran ambos, componentes y compuestos, de forma simultánea.

No sólo en relación a la indexación de sintagmas sino también en la discriminación de sentidos resulta difícil justificar un procesamiento lingüístico sofisticado. La selección explícita del sentido de las palabras tampoco tiene por qué mejorar la recuperación de información (Krovetz 1992) y para esta tarea, pueden aplicarse técnicas estadísticas incluso con mejores resultados (Schütze 1995).

La indexación conceptual basada en synsets de WordNet propuesta por (Gonzalo 1998b) resulta prometedora. Una vez determinado el sentido de las palabras en el texto, se sustituyen por una representación que resulta independiente del idioma, habilitando de forma directa no sólo la consideración de sinónimos, sino también la posibilidad de que la recuperación sea multilingüe utilizando para ello los synsets de EuroWordNet.

Para estudiar la viabilidad de un sistema real en el que la desambiguación del sentido de las palabras debe realizarse de forma automática, es necesario determinar en que grado el modelo de indexación conceptual basado en synsets es sensible a la introducción de errores en la desambiguación del sentido de las palabras.

Si bien (Sanderson 1994;Sanderson 2000) realiza el estudio de cómo afecta a la recuperación la introducción de errores en la desambiguación de pseudo-palabras, es necesario comprobar si la utilización de synsets sigue el mismo patrón o si por el contrario no es necesario alcanzar el 90% de precisión en la desambiguación para mejorar la recuperación.

2.5.2 Exploración de términos

La mayoría de los experimentos muestran que en los modelos tradicionales de recuperación de documentos no se justifica ni se necesita una indexación motivada lingüísticamente (Sparck Jones 1999).

Es en tareas de más alto nivel en el acceso, discriminación, descripción y selección de información donde las técnicas lingüísticas pueden aportar elementos de mejora.

Para ello es necesario retomar la perspectiva de acceso a la información más allá de la mera recuperación de documentos.

Una de las áreas donde las técnicas lingüísticas parecen tener cabida es en la exploración de sintagmas como vía de acceso a la información. El ruido producido por la ambigüedad de las palabras al expandir o traducirse a otro idioma las palabras de la consulta, se reduce en gran medida cuando se consideran las restricciones que se imponen entre sí las componentes de los sintagmas.

Sin embargo, los sistemas actuales de exploración de términos tampoco pueden abordar los problemas de variación morfosintáctica, semántica y translingüe.

Capítulo 3

Experimentos en ambigüedad léxica e indexación

Uno de los problemas más importantes que subyace al lenguaje natural y que afecta de forma directa a la recuperación de información es la ambigüedad léxica. La ambigüedad de las expresiones lingüísticas es inherente a todos los niveles léxicos: morfológico, sintáctico y semántico. Esto lleva consigo una inevitable pérdida de información y una introducción de ruido al realizar inferencias lingüísticas. El problema se hace presente de forma especial cuando se tratan los problemas de variación terminológica y de multilingüismo al tener que seleccionar los elementos de expansión o traducción tanto en las consultas como en los documentos, ya sea a otro idioma como a una representación conceptual.

Resulta difícil determinar como afecta a la Recuperación de Información la ambigüedad léxica en cada uno de los niveles léxicos. La consideración explícita de la ambigüedad léxica requiere un procesamiento lingüístico: etiquetado morfosintáctico para determinar la categoría gramatical y desambiguación del sentido de las palabras (WSD, *Word Sense Disambiguation*). Este procesamiento lingüístico no está exento de errores e introduce una variable difícil de cuantificar a la hora de evaluar los efectos de un procesamiento lingüístico en la Recuperación de Información. Estos errores pueden perjudicar la recuperación compensando los posibles beneficios en el uso de técnicas lingüísticas. Resulta necesario esclarecer en que medida los pobres resultados de recuperación utilizando técnicas lingüísticas en la indexación se deben a que éstas no son adecuadas o a los errores que introduce el procesamiento automático.

Los siguientes experimentos abordan esta cuestión y estudian la viabilidad de un sistema basado en indexación conceptual. En primer lugar se describirá la colección utilizada para los experimentos (IR-SEMCOR). Se trata de una colección con la particularidad de que tanto consultas como documentos están etiquetados manualmente en todos los niveles léxicos: categoría gramatical, lema, sentido y synset en WordNet 1.5. Debido a que el etiquetado manual tiene una tasa de error muy baja (en torno al 10% en SEMCOR), la colección permite dilucidar cómo afectan los errores de procesamiento lingüístico automático en la recuperación. Tras la descripción de la colección (3.1), se explican los experimentos realizados con este fin en el nivel morfosintáctico (3.2), (3.3) y semántico (3.4), (3.5), (3.6), (3.7).

Por último, se presenta un trabajo colectivo (3.8), el motor de búsqueda ITEM, basado en indexación conceptual sobre synsets de WordNet que servirá para evaluar cualitativamente este tipo de indexación.

3.1 La colección de prueba IR-SEMCOR

La colección SEMCOR (Miller 1993) consta de un subconjunto de unos 100 documentos pertenecientes al BROWN CORPUS que han sido etiquetados con los sentidos de WordNet. Esta colección tiene unos 2.4Mb de texto (22Mb incluyendo las anotaciones). Es una colección heterogénea sobre política, deportes, música, cine, filosofía, extractos de novelas, textos científicos, etc.

Para convertir SEMCOR en la colección de prueba para recuperación de información (IR-SEMCOR), se realizó el siguiente trabajo:

1. División de los textos de SEMCOR 1.5 en documentos coherentes para Recuperación de Información. Algunos de los textos originales contenían partes que trataban de temas diferentes y fue necesario separarlas para formar los documentos de la colección. Obtuvimos 171 documentos con una longitud media de 1331 palabras. Posteriormente, los documentos de SEMCOR 1.6 fueron añadidos a la colección sin modificaciones (aparte del mapping de sentidos de WordNet 1.6 a WordNet 1.5) hasta completar un total de 254 documentos.

2. Para cada uno de los primeros 171 documentos, se escribió un breve resumen de entre 4 y 50 palabras (22 palabras por resumen en media). Cada resumen es una explicación coherente del contenido del texto y no una mera lista de palabras clave.

3. Finalmente, cada uno de los 171 resúmenes fue etiquetado manualmente con sentidos de WordNet 1.5. Cuando una palabra o término multipalabra no estaba presente en las bases de datos se dejaba el término original sin desambiguar. En general, estos términos correspondían a nombres propios de grupo (v.g. Fulton_county_Grand_Jury), persona (v.g. Cervantes) o localidades (v.g. Fulton)

Por ejemplo, la primera consulta de IR-SEMCOR es:

The Fulton County Grand Jury investigates possible irregularities Atlanta's primary election

cuya versión en lemas, sentidos y synsets etiquetados manualmente son respectivamente las siguientes:

The Fulton_County_Grand_Jury investigate possible irregularity in atlanta primary_election

Fulton_County_Grand_Jury investigate%2:32:01:: possible%3:00:04:: irregularity%1:04:00:: atlanta%1:15:00 primary_election%1:04:00::

Fulton_County_Grand_Jury v00441414 a00036893 n00412042 n5608324 n00103176

De cara al proceso de recuperación, también se desarrollaron listas de "stop-senses" y "stop-synsets" de forma automática a partir de una lista de "stopwords" para el inglés.

En los experimentos realizados en (Gonzalo 1998b;Gonzalo 1999b) los resúmenes fueron utilizados como consultas en las que cada una de ellas debía recuperar exactamente un documento, aquel que resumía. En esta ocasión, y con el fin de conseguir unos juicios de relevancia más estándar, se ha realizado el siguiente presupuesto: si un texto original de SEMCOR fue dividido en n documentos de la colección de recuperación, entonces el resumen de cada uno de estos documentos debe recuperar todos los documentos pertenecientes al mismo texto original. En esta ocasión, todos los resúmenes de textos originales que no fueron divididos no se han considerado.

De esta forma, esta versión de IR-SEMCOR tiene 82 consultas etiquetadas con sentidos de WordNet 1.5, con una media de 6.8 documentos relevantes por consulta.

Con el fin de evaluar la plausibilidad de este conjunto de juicios de relevancia, se ha producido de forma aleatoria un juego alternativo de juicios de relevancia. Estos juicios se muestran como línea base de comparación en los experimentos que se describen a continuación.

3.2 Ambigüedad morfosintáctica en Recuperación de Información

Para tratar de discernir en qué grado la pérdida de efectividad en la recuperación se debe a los errores de etiquetado de categoría gramatical, el siguiente experimento reproduce el trabajo de (Krovetz 1997), pero aprovechando que IR-SEMCOR permite comparar el etiquetado manual con el etiquetado automático del *Brill POS Tagger* (Brill 1992).

3.2.1 Definición del experimento

El objetivo es comparar la efectividad de la recuperación con tres indexaciones diferentes:

1. Indexación con texto plano.
2. Indexación ligando a las palabras su categoría gramatical asignada manualmente.
3. Indexación ligando a las palabras su categoría gramatical asignada automáticamente.

Las tres indexaciones así como la recuperación de documentos se realizan con el mismo motor de búsqueda (INQUERY). Estas comparaciones permitirán determinar el comportamiento en la colección IR-SEMCOR del etiquetado automático respecto a la indexación sin etiquetado o con etiquetado manual.

3.2.2 Realización del experimento y resultados

La *Figura 3-1* muestra la efectividad de la recuperación en términos de *precisión* y *recall*. La figura muestra que la recuperación tras un etiquetado de categoría no se ve significativamente alterado respecto a la recuperación sin etiquetado, y que ello resulta independientemente de que se haya realizado un etiquetado de forma manual o automática.

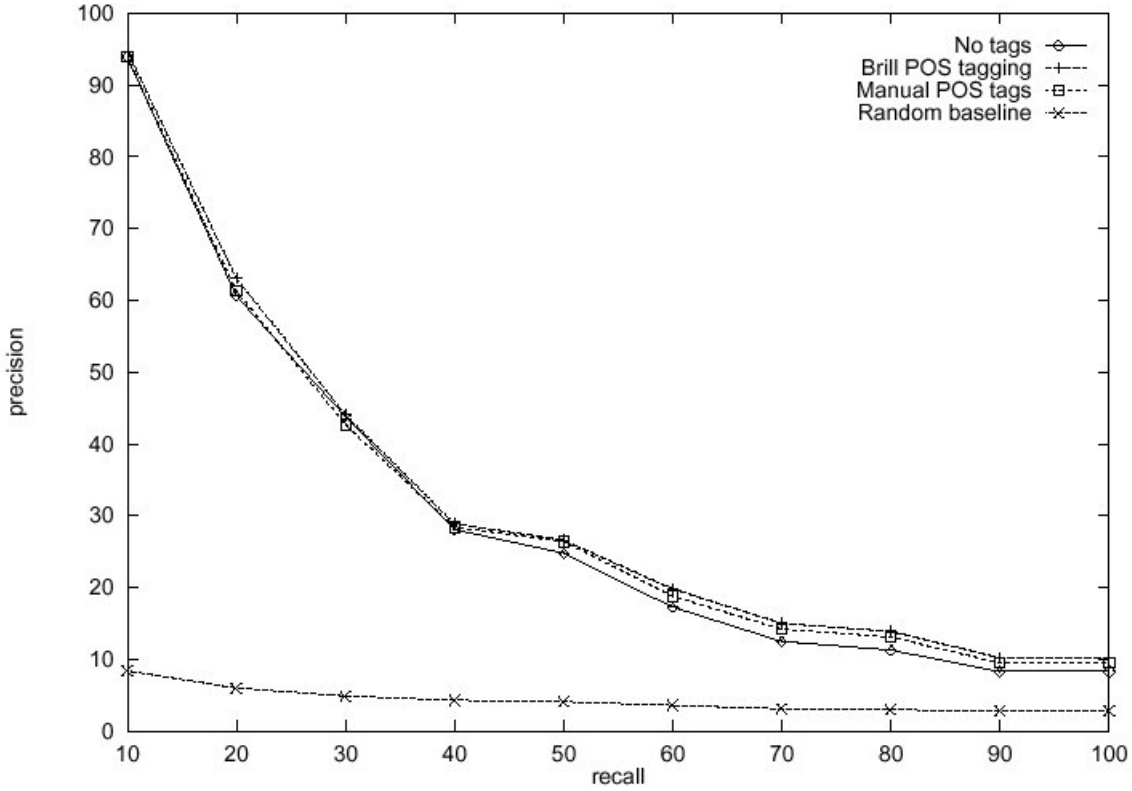


Figura 3-1. Efectos del etiquetado en la Recuperación de Información

Una exploración más detallada muestra ejemplos concretos en los que se produce un perjuicio, un beneficio o una compensación de los errores en el etiquetado en cada tipo de indexación.

3.2.2.1 Etiquetado manual vs. texto plano

La exploración cualitativa de los resultados permite comprobar que hay varios efectos que se compensan. En los casos en que el ajuste de términos coincide en categoría, los documentos son puntuados mejor en el ranking debido a que el etiquetado reduce el número de documentos recuperados. Sin embargo, la distinción de categorías disminuye la cobertura respecto al stemming de texto plano.

Por ejemplo, en la colección sin etiquetado, una consulta que contiene *"talented baseball player*" ajusta con *"is one of the top talents of the time"*, debido a que el stemming reduce *talent* y *talented* a la misma forma sin importar la categoría gramatical. Sin embargo, tras el etiquetado se intenta ajustar sin éxito *talent/ADJ*

con *talent/N*. Otro ejemplo es *"skilled diplomat of an Asian Country"* respecto a *"diplomatic policy"*, donde *diplomat/N* no ajusta con *diplomat/ADJ*.

Estos dos efectos parecen compensarse para producir curvas similares de precisión/cobertura.

3.2.2.2 Etiquetado manual vs. etiquetado automático

Debido a que el etiquetado en un sistema real deberá realizarse de forma automática, es conveniente tratar de evaluar en que grado los errores de etiquetado automático afectan a la recuperación.

Aunque el etiquetado automático produce más errores que el etiquetado manual (que tampoco está exento de ellos), curiosamente no todos los errores afectan negativamente a la recuperación. Por ejemplo, para una consulta que contiene *"summer/N shoe/N design/N",* el etiquetador de Brill produce *"summer/N shoe/N design/V".* Un documento que contiene "*Italian designed sandals*" cuyo etiquetado manual es "*Italian/ADJ designed/ADJ sandals/N"* resulta etiquetado por el Brill como "*Italian/ADJ designed/V sandals/N".* En este caso, mediante el etiquetado manual *design* como nombre en la consulta y como adjetivo en el documento, no ajustan. Sin embargo, los errores del etiquetado automático sí permiten que *design* como verbo tanto en la consulta como en el documento ajusten tras el stemming.

Esta compensación de errores hace que la recuperación con etiquetado manual o automático también se comporten estadísticamente de una forma similar.

3.2.3 Conclusiones

Así pues, el etiquetado no parece una estrategia que se justifique con fines de Recuperación de Información. El coste de etiquetado es muy elevado y sin embargo la recuperación no mejora significativamente. Sin embargo, el etiquetado es un paso intermedio hacia otro tipo de indexación (e.g. basada en sentidos) que sí puede producir alguna mejora.

3.3 Indexación de sintagmas en Recuperación de Información

Casi la mitad de los términos contenidos en WordNet corresponden a colocaciones o términos multipalabra (más de 55.000). Todas estas colocaciones están etiquetadas

como tal en la colección IR-SEMCOR pero, además, IR-SEMCOR tiene etiquetados sintagmas nominales relativos a personas, grupos, localidades, instituciones, etc. que no están contemplados en WordNet y en general se corresponden con nombres propios (v.g. Drew Centennial Church o Mayor-nominate Ivan Allen Jr.).

3.3.1 Definición del experimento

Aprovechando que la colección IR-SEMCOR tiene todos estos sintagmas etiquetados de forma manual, se ha diseñado un experimento para comprobar sin errores de procesamiento automático, si la indexación de sintagmas permite mejorar o no la recuperación. El experimento compara tres estrategias de recuperación diferentes:

1. Texto plano tanto en consultas como documentos, sin considerar información alguna sobre sintagmas.

2. Consideración como unidades de indexación, de los sintagmas etiquetados manualmente tanto en consultas como en documentos. Por ejemplo, *New_York* es un término que no podrá relacionarse con *new* o *York*, lo que parece beneficioso. Por otro lado, *Drew_Centennial_Church* será un termino no relacionado con *church* lo que puede provocar un aumento de precisión, pero también una pérdida de documentos relevantes.

3. Texto plano en documentos, pero aprovechando el operador #*phrase* del motor de búsqueda utilizado (INQUERY), para los sintagmas recogidos en las consultas. Por ejemplo, "*meeting of the National_Fotball_League",* se expresa como

   #sum(meeting #phrase(National Football League))

   en el lenguaje de consulta. El operador #*phrase* asigna un crédito parcial a las componentes del sintagma, premiando los casos de co-ocurrencia.

De esta forma se comparan los efectos de la recuperación cuando se toman los sintagmas lexicalizados de WordNet como unidades de indexación, y la recuperación cuando además del sintagma se otorga un crédito parcial a las componentes del sintagma.

3.3.2 Realización del experimento y resultados

Los resultados del experimento pueden observarse en la *Figura 3-2*. La gráfica muestra que la indexación con sintagmas se comporta ligeramente peor que la recuperación estándar sobre texto plano y que la recuperación utilizando el operador #*phrase* es similar a la recuperación sin sintagmas.

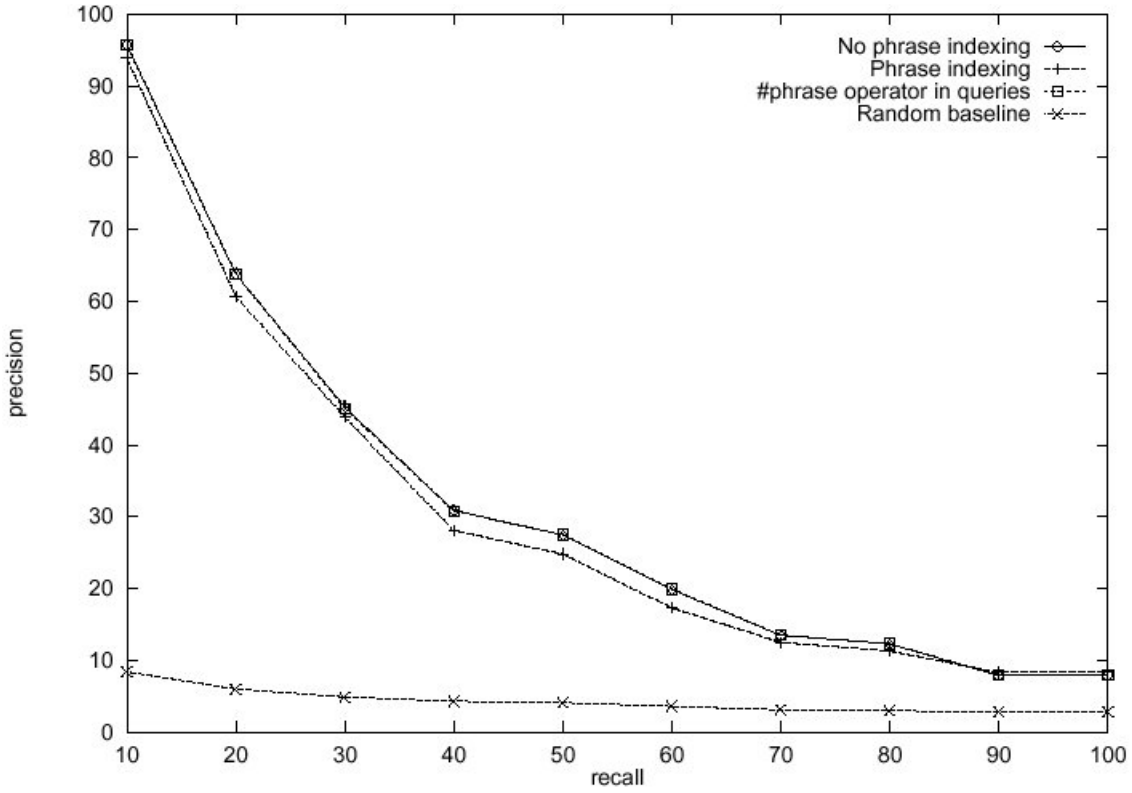


Figura 3-2. Indexación de sintagmas en IR-SEMCOR

El estudio cualitativo revela más información acerca de lo que está pasando:

- En algunos casos, la restricción de adyacencia que impone el sintagma es demasiado fuerte. Así, una consulta que contiene *"candidate in governor's_race"* no se ajusta con *"opened his race for governor"*, lo cual muestra no sólo que es conveniente considerar las componentes del sintagma, sino que también lo es relajar la restricción de adyacencia y considerar una ventana de más de una palabra. Muchos sistemas utilizan esta idea de proximidad para relajar la restricción de adyacencia. En estos sistemas se da mayor peso a las palabras más próximas dentro de una determinada ventana. La proximidad física entre palabras tiene un valor

semántico, aunque no se sabe muy bien cómo incorporarlo para mejorar el ranking de documentos.

- Cuanto más largas son las consultas, más aporta a la recuperación la consideración de sintagmas. El mayor número de palabras en la consulta hace crecer el número de documentos recuperados y la consideración de sintagmas permite aflorar aquellos que son más relevantes. En estos casos, la restricción de adyacencia resulta conveniente y se comporta mejor que el operador de proximidad #*phrase* que permite ajustes parciales erróneos.

- La restricción de adyacencia que imponen los sintagmas aumenta la precisión de las búsquedas pero a costa de una disminución drástica en la cobertura. La consideración de sintagmas puede ser conveniente cuando las búsquedas deben ser precisas a causa de dominios muy específicos o de elementos de búsqueda muy concretos. Esto no ocurre en nuestra colección debido a que los juicios de relevancia en IR-SEMCOR se asignaron en función de los fragmentos de un mismo texto original en SEMCOR. En nuestra colección los juicios de relevancia dependen de que los temas de búsqueda estén relacionados, más que de encontrar un elemento concreto de búsqueda. Por ejemplo, la consideración de sintagmas no permite relacionar *"story of a famous strip cartoonist"* con un documento que contiene *"detective_story"* lo cual parece adecuado si se interpreta estrictamente la consulta pero que, sin embargo, en IR-SEMCOR pertenece al mismo texto original y, por tanto, se consideran relacionados en los juicios de relevancia. Lo mismo ocurre con la consulta *"The board_of_regents of Paris_Junior_College has named the school's new president", que no se relaciona con "Junior or Senior High School Teaching Certificate"*. Esto parece apropiado, pero no lo es según los juicios de relevancia de IR-SEMCOR.

3.3.3 Conclusiones

El estudio de la casuística que afecta a la recuperación positiva o negativamente revela, ante todo, que hay una componente composicional muy importante en el significado de un sintagma pero que esto no ocurre siempre. Cuando las componentes de un sintagma no pierden su significado por el hecho de pertenecer al mismo, entonces la consideración del sintagma como una sola unidad de indexación resulta contraproducente.

Esto sugiere la necesidad de plantear una tipología de sintagmas y a considerarlos de manera diferente en Recuperación de Información dependiendo de cómo se ve afectado el significado de las componentes por el significado del compuesto.

3.4 Distinción de compuestos léxicos en IR

En este apartado se proponen los criterios para realizar una clasificación de los compuestos léxicos de WordNet considerando sus componentes, y se estudia cómo afecta a la recuperación esta clasificación.

3.4.1 Tipos de compuestos léxicos

Cabe realizar la siguiente distinción semántica de compuestos nominales en inglés:

3.4.1.1 Compuesto Endocéntrico.

Este tipo de compuesto hereda las propiedades sintácticas (género, número, etc.) de una de las componentes (en inglés, generalmente la segunda palabra componente). Semánticamente, el compuesto es un hipónimo del segundo elemento, siendo el primero un modificador. Por ejemplo, *armchair* que es un tipo de *chair* o *"toothed whale"* que es un tipo de *whale*.

3.4.1.2 Compuesto Exocéntrico

Este tipo de compuesto también hereda las propiedades sintácticas de su última componente, pero semánticamente es un hipónimo de un elemento no expresado en el compuesto. Por ejemplo, *hatchback* es un tipo de *car*, o "*mentally retarded*" es un tipo de *people.*

3.4.1.3 Compuesto Aposicional

Este tipo de compuesto también conserva el último elemento como principal de una forma sintáctica pero, semánticamente, el compuesto hereda propiedades de ambas componentes. Por ejemplo, *"folk song"* que es un tipo tanto de *music* como de *folk*. O como, por ejemplo, *girlfriend.* En este caso no se pierde mucha información si estos compuestos se tratan como compuestos endocéntricos ya que no hay mucha

diferencia entre una representación semántica de *"both, girl and friend"* o *sólo "a friend who is a girl"*.

3.4.1.4 Compuesto Copulativo (o compuesto Dvandva)

En estos compuestos no siempre se distingue cuál de las componentes es la principal sintáctica o semánticamente. Son entidades separadas que se combinan para referenciar a una única entidad como, por ejemplo, *pantyhose* (*pantis, medias*). Salvo este ejemplo, en general se trata de nombres propios.

3.4.2 Propuesta de clasificación automática de compuestos léxicos mediante WordNet

En WordNet 1.5 hay más de 56.000 compuestos multipalabra. De ellos, tan sólo 2860 son polisémicos (5%). Estos compuestos polisémicos han sido excluidos del proceso de clasificación que se describe más adelante.

Los compuestos copulativos no están en WordNet. El resto de tipos de compuestos sí están presentes en WordNet y las relaciones de hiperonimia/hiponimia y sinonimia entre compuesto y componentes van a permitir detectarlos.

El proceso de clasificación tiene que tener en cuenta que una componente de un compuesto multipalabra puede ser a su vez otro compuesto multipalabra. También debe tener en cuenta que, en WordNet, hay ocasiones en las que una componente es sinónimo del compuesto y, por tanto, la clasificación de compuestos no debe considerar únicamente hiperónimos, sino también sinónimos.

3.4.2.1 Clasificación de compuestos copulativos

Los *compuestos copulativos (o Dvandva)* son compuestos que no se encuentran en WordNet. Su detección con fines de recuperación de información es muy interesante, pero requiere un proceso de extracción de entidades. En SEMCOR están etiquetados con los *valores group, location, person, other* para la etiqueta *rdf*.

En la colección IR-SEMCOR hay 8.984 compuestos multipalabra de este tipo, de los cuales 4.153 son diferentes. Esto supone un porcentaje de términos bajo, aunque su poder discriminativo es elevado.

3.4.2.2 Clasificación de compuestos endocéntricos

La forma de distinguir estos compuestos es comprobando que una de sus componentes es hiperónimo del compuesto. Puede ocurrir que se trate del hiperónimo inmediatamente superior, que se trate de otro más alto en la jerarquía, incluso que se den ambos casos simultáneamente como ilustra el ejemplo:

```
1 sense of atlantic bottlenose dolphin

Sense 1
Atlantic bottlenose dolphin, Tursiops truncatus
      => bottlenose dolphin, bottle-nosed dolphin, bottlenose
         => dolphin
            => toothed whale
               => whale
                  => cetacean, cetacean mammal, blower
                     => aquatic mammal
```

Sin embargo, para considerar a estos compuestos como endocéntricos, los diferentes hiperónimos deben pertenecer a la misma rama de la jerarquía. En otro caso se trataría de compuestos aposicionales como se verá más adelante. Otros ejemplos de compuestos endocéntricos son los siguientes:

```
1 sense of primary election

Sense 1
primary, primary election
      => election
         => vote
            => group action
               => act, human action, human activity
```

```
1 sense of grand jury

Sense 1
grand jury
      => jury
         => body
            => gathering, assemblage
               => social group
                  => group, grouping
```

```
1 sense of purchasing department

Sense 1
purchasing department
      => business department
         => department
            => division
               => administrative unit
                  => unit
                     => organization
                        => social group
                           => group, grouping
```

```
1 sense of alloy steel

Sense 1
alloy steel
      => steel
         => alloy
            => mixture
               => substance, matter
                  => object, inanimate object, physical object
                     => entity
            => metallic element, metal
               => chemical element, element
                  => substance, matter
                     => object, inanimate object, physical object
                        => entity
```

En algunos casos, los compuestos no tienen ninguna componente hiperónima pero, sin embargo, existe alguna componente que es sinónimo del compuesto. Estos casos también se van a clasificar como compuestos endocéntricos. Ejemplos:

```
1 sense of bank account

Sense 1
account, bank account
      => fund, monetary fund
         => money
            => medium of exchange, monetary system
               => asset
                  => possession
```

```
1 sense of abraham lincoln

Sense 1
Lincoln, Abraham Lincoln
      => lawyer, attorney
         => professional, professional person
            => adult
               => person, individual, someone, mortal, human, soul
                  => life form, organism, being, living thing
                     => entity
                  => causal agent, cause, causal agency
                     => entity
      => President of the United States, President, Chief Executive
         => head of state, chief of state
            => leader
               => person, individual, someone, mortal, human, soul
                  => life form, organism, being, living thing
                     => entity
                  => causal agent, cause, causal agency
                     => entity
```

También se da el caso en que una componente es sinónimo e hiperónimo del compuesto simultáneamente. Aunque en compuestos verbales este tipo es un poco más común, en compuestos nominales sólo hay dos ejemplos:

```
1 sense of electric drill

Sense 1
drill, electric drill
      => power drill
         => power tool
            => machine
               => device
                  => instrumentality, instrumentation
                     => artifact, artefact
                        => object, inanimate object, physical object
                           => entity
         => drill
            => tool
               => implement
                  => instrumentality, instrumentation
                     => artifact, artefact
                        => object, inanimate object, physical object
                           => entity
```

```
1 sense of kentucky yellowwood

Sense 1
Kentucky yellowwood, gopherwood, Cladrastis lutea, Cladrastis kentukea, yellowwood
      => angiospermous yellowwood
         => yellowwood, yellowwood tree
            => tree
               => woody plant, ligneous plant
                  => vascular plant, tracheophyte
                     => plant, flora, plant life
                        => life form, organism, being, living thing
                           => entity
```

3.4.2.3 Clasificación de compuestos aposicionales

La forma de detectar estos compuestos es comprobar que más de una componente diferente es hiperónimo del compuesto pero en jerarquías también diferentes (herencia múltiple). Ejemplos:

```
1 sense of aspirin powder

Sense 1
aspirin powder, headache powder
      => aspirin, acetylsalicylic acid, Bayer, Empirin
         => analgesic, anodyne, painkiller, pain pill
            => medicine, medication, medicament, medicinal drug
               => drug
                  => artifact, artefact
                     => object, inanimate object, physical object
                        => entity
      => powder
         => toiletry, toilet article, toiletries
            => instrumentality, instrumentation
               => artifact, artefact
                  => object, inanimate object, physical object
                     => entity
         => medicine, medication, medicament, medicinal drug
            => drug
               => artifact, artefact
                  => object, inanimate object, physical object
                     => entity
```

```
1 sense of folk song

Sense 1
folk song, folk ballad
      => folk music, ethnic music, folk
         => music
            => art, fine art
               => creation
                  => artifact, artefact
                     => object, inanimate object, physical object
                        => entity
      => song
         => musical composition, opus, composition, piece, piece of music
            => music
               => art, fine art
                  => creation
                     => artifact, artefact
                        => object, inanimate object, physical object
                           => entity
```

```
1 sense of compact disc read-only memory

Sense 1
CD-ROM, compact disc read-only memory
      => compact disc, compact disk, CD
         => recording
            => memory device, storage device
               => device
                  => instrumentality, instrumentation
                     => artifact, artefact
                        => object, inanimate object, physical object
                           => entity
      => read-only memory, ROM, read-only storage, fixed storage
         => memory, storage, store, memory board
            => memory device, storage device
               => device
                  => instrumentality, instrumentation
                     => artifact, artefact
                        => object, inanimate object, physical object
                           => entity
```

En los compuestos aposicionales se van a incluir aquellos en los que una componente es sinónimo del compuesto y otra componente distinta es un hiperónimo. Por ejemplo,

1 sense of 1st-class mail

Sense 1
first-class, 1st-class, first-class mail, 1st-class mail, priority mail
=> mail
=> message
=> communication
=> social relation
=> relation
=> abstraction

1 sense of abductor muscle

Sense 1
abductor, abductor muscle
=> skeletal muscle
=> muscle, musculus
=> contractile organ
=> organ
=> body part
=> part, piece
=> entity

1 sense of abrasive material

Sense 1
abrasive, abradant, abrasive material
=> material, stuff
=> substance, matter
=> object, inanimate object, physical object
=> entity

1 sense of african green monkey

Sense 1
green monkey, **African green monkey**, Cercopithecus aethiops sabaeus
=> guenon, guenon monkey
=> Old World monkey
=> **monkey**
=> primate
=> placental mammal, eutherian, eutherian mammal
=> mammal
=> vertebrate, craniate
=> chordate
=> animal, animate being, beast, brute, creature, fauna
=> life form, organism, being, living thing
=> entity

```
1 sense of future perfect tense

Sense 1
future perfect, future perfect tense
      => perfect, perfect tense
         => tense
            => grammatical category, syntactic category
               => class, category, family
                  => collection, aggregation, accumulation, assemblage
                     => group, grouping
```

3.4.2.4 Clasificación de compuestos exocéntricos

Para detectar estos compuestos hay que comprobar que no tiene ninguna componente que sea sinónimo o hiperónimo del compuesto. Ejemplos:

```
1 sense of fisher cat

Sense 1
fisher, pekan, fisher cat, black cat, Martes pennanti
      => marten, marten cat
         => musteline mammal, mustelid, musteline
            => carnivore
               => placental mammal, eutherian, eutherian mammal
                  => mammal
                     => vertebrate, craniate
                        => chordate
                           => animal, animate being, beast, brute, creature, fauna
                              => life form, organism, being, living thing
                                 => entity
```

```
1 sense of man and wife

Sense 1
marriage, married couple, man and wife
      => family, family unit
         => kin, kin group, kinship group, kindred, clan, tribe
            => social group
               => group, grouping
```

```
1 sense of mentally retarded

Sense 1
mentally retarded
      => people
         => group, grouping
```

```
1 sense of lieutenant governor

Sense 1
lieutenant governor
      => elected official
         => official, functionary
            => worker
               => person, individual, someone, mortal, human, soul
                  => life form, organism, being, living thing
                     => entity
                  => causal agent, cause, causal agency
                     => entity
```

```
1 sense of per cent

Sense 1
percentage, percent, per cent, pct
      => proportion, proportionality
         => quotient
            => ratio
               => magnitude relation
                  => relation
                     => abstraction
```

3.4.3 Propuesta de distinción de compuestos léxicos en Recuperación de Información

Los lenguajes de consulta incluyen operadores de proximidad y de adyacencia que permiten que en los documentos no haya que detectar ni unir sintagmas. Así pues, es en la consulta donde los compuestos serán representados de formas distintas en función de su tipología.

En esta propuesta se van a distinguir tres grupos de compuestos de cara a la Recuperación de Información:

1. Compuestos copulativos.
2. Compuestos exocéntricos.
3. Compuestos endocéntricos y aposicionales.

3.4.3.1 Distinción de compuestos copulativos en IR

No hay criterios claros para decidir si las componentes de un compuesto copulativo deben conservarse unidas o separadas. En principio conviene considerar ambas posibilidades simultáneamente, es decir, mantener unido el compuesto multipalabra y añadir algunas componentes. Un posible criterio es no añadir componentes muy polisémicas y/o muy genéricas (altas en la jerarquía).

Las entidades representadas por los compuestos copulativos tienen el problema de su detección (no están en WordNet) y de que pueden presentarse varias formas distintas para el mismo referente. La dificultad proviene de que dichas formas no son previsibles. Son necesarias bases de datos con las posibles referencias para un mismo referente: Clinton/Bill Clinton, Gorbachov/Gorbachev/Gorbi, etc. Se trata de un problema de *Reconocimiento de Entidades y Extracción de Informaci*ón en el que no entraremos aquí.

3.4.3.2 Distinción de compuestos exocéntricos en IR

En este caso, puesto que las componentes pierden su significado, resulta claro que las componentes del compuesto no deben considerarse por separado. La exploración de los índices de WordNet 1.5 permite identificar 19.284 compuestos exocéntricos incluidas todas las categorías. Esto supone que el 34% de los compuestos multipalabra de WordNet son exocéntricos.

3.4.3.3 Distinción de compuestos endocéntricos y aposicionales en IR

Los compuestos aposicionales pueden considerarse un caso particular de los compuestos endocéntricos en los que son varias y no una las componentes hiperónimas o sinónimas del compuesto.

En este caso, las componentes no pierden su significado, sino que modifican un sentido nuclear y, por tanto, es preferible mantenerlas separadas. No hay criterios claros ni determinantes para decidir si la componente nuclear debe pesarse más o menos que las demás. Si queremos centrar la búsqueda en el topic general de la consulta, parece conveniente pesar más la componente nuclear del compuesto (la componente hiperónima).

En el caso de que la componente hiperónimo no sea del nivel inmediatamente superior de la jerarquía de WordNet, parece apropiado añadir a la consulta, a modo de expansión, los términos de los synsets de niveles intermedios. Por ejemplo, en el

caso de *"abstract artist"*, *artist* es hiperónimo de segundo nivel y podría añadirse *painter*:

```
1 sense of abstract artist

Sense 1
abstractionist, abstract artist
      => painter
         => artist, creative person
            => creator
               => person, individual, someone, mortal, human, soul
                  => life form, organism, being, living thing
                     => entity
                  => causal agent, cause, causal agency
                     => entity
```

3.4.4 Definición del experimento

El experimento tiene como objetivo comparar la recuperación en términos de precisión y cobertura, cuando se consideran compuestos léxicos. Para ello se ha utilizado la colección de prueba OHSUMED que tiene 380Mb de documentos y 101 consultas en el dominio médico. Debido a que la clasificación de compuestos expuesta en el apartado anterior se realiza sobre WordNet, su utilidad en Recuperación de Información depende de lo bien que WordNet cubra el dominio de búsqueda. En este caso, la colección OHSUMED resulta apropiada para el experimento porque las subjerarquías de WordNet relativas al dominio médico son bastante ricas y, por tanto, se espera que la recuperación se vea afectada por la distinción de compuestos.

El motor de búsqueda empleado ha sido INQUERY, las colecciones se han indexado en formato texto original y sólo se han procesado las consultas de acuerdo con el tratamiento descrito anteriormente. Los experimentos que se van a comparar son los siguientes:

1. Sin compuestos. Las consultas no se han procesado en ningún sentido salvo para adecuarlas al lenguaje de consulta del motor de búsqueda.

2. Adyacencia. A todos los compuestos detectados en las consultas se les ha impuesto la restricción de que en el texto debe encontrarse exactamente la misma secuencia, sin posibilidad de considerar las componentes aisladas (operador *#ws, window size*, igual al número de palabras del compuesto).

3. Proximidad. En este caso, en lugar de exigir la adyacencia de las palabras del compuesto, se pide que aparezcan en un entorno próximo, pero además otorgando un crédito parcial a la ocurrencia aislada de las componentes en el texto (operador #*phrase*).

4. Restricción de adyacencia para compuestos exocéntricos y de proximidad para el resto de compuestos. A los compuestos exocéntricos se les impone la restricción de adyacencia (operador #*ws* con tamaño igual al número de componentes), mientras que al resto de compuestos en las consultas se les aplica el operador de proximidad (#*phrase*).

5. Restricción de adyacencia sólo para compuestos exocéntricos.

6. Restricción de adyacencia con sobrepeso sólo para compuestos exocéntricos. En este caso, a los compuestos exocéntricos, además de restringir la ventana al número de componentes, se les aplica un sobrepeso (operador #+).

3.4.5 Realización del experimento y resultados

La *Tabla 3-1* muestra los resultados obtenidos en términos de precisión/recall:

1. Sin compuestos. La precisión media en los 10 puntos de recall es del 19.2%.

2. Adyacencia. En este experimento, la precisión media baja a 15.8%, lo que supone una pérdida del 17.7%.

3. Proximidad. La precisión media sube al 18.4% pero no llega a la precisión obtenida sin la consideración de compuestos (19.2%).

4. Restricción de adyacencia para compuestos exocéntricos y de proximidad para el resto de compuestos. La precisión media prácticamente coincide con la anterior, siendo del 18.3%.

5. Restricción de adyacencia sólo para compuestos exocéntricos. En este caso la precisión media sube a los niveles de recuperación sin compuestos, siendo del 19.3%, apenas una décima por encima.

6. Restricción de adyacencia con sobrepeso sólo para compuestos exocéntricos. La precisión media apenas sube una décima más hasta llegar al 19.4%.

	Precisión (101 consultas)					
Recall	Sin compuestos	Adyacencia	Proximidad	Adyacencia exocentric. Proximidad resto	Adyacencia exocentric.	Adyacencia exocentric. con más peso
10	44.4	40.9	43.3	43.1	44.5	44.5
20	35.7	32.3	35.6	35.6	37.3	37.4
30	29.0	23.4	27.5	27.3	29.3	29.4
40	23.4	19.0	22.1	22.0	23.2	23.3
50	19.7	15.1	18.6	18.6	19.9	19.9
60	13.8	11.1	12.7	12.7	13.6	13.7
70	10.4	7.3	9.5	9.5	10.2	10.2
80	7.7	5.1	7.0	7.1	7.4	7.4
90	4.9	2.7	4.4	4.3	4.7	4.7
100	3.0	1.5	2.9	2.9	3.0	3.0
Media	19.2	15.8	18.4	18.3	19.3	19.4

Tabla 3-1 Distinción de compuestos en Recuperación de Información

3.4.6 Conclusiones

En todos los casos en los que se consideran sintagmas se pierde efectividad en la recuperación, salvo cuando se consideran únicamente compuestos exocéntricos. Este comportamiento respecto a los compuestos exocéntricos es lógico puesto que en este caso las componentes no mantienen un significado parcial del compuesto y por tanto la consideración de las componentes por separado conduce a resultados incorrectos.

Al igual que (Fagan 1989), hay que destacar que el número de compuestos léxicos en las consultas resulta muy reducido, lo que impide que las diferencias entre los experimentos no sean determinantes. El hecho de que un sobrepeso sobre los compuestos exocéntricos eleve algo la precisión media parece indicar que resulta conveniente su detección y consideración. Sin embargo, las diferencias en la recuperación son casi inapreciables y no justifican el coste computacional que se añade al procesamiento de las consultas.

3.5 *Synsets de Variantes Monosémicas*

Si bien las palabras más frecuentes en cualquier colección son las más polisémicas, alrededor del 20% de las palabras de un texto son monosémicas. En WordNet 1.5 el 80% de los términos contenidos en la base de léxica son términos monosémicos. Esto, significa que hay un gran número de palabras cuya desambiguación es trivial y, por tanto, su traducción a synsets también lo es. Sin embargo, no es así su consideración con fines de indexación.

3.5.1 Definición de Synset de Variantes Monosémicas

El interés de una indexación conceptual es que un documento puede ser recuperado a partir de términos sinónimos a los que contiene. Sin embargo, que un término sea monosémico no implica que sus sinónimos también lo sean. Así aunque sea trivial desambiguar un término monosémico, no lo es desambiguar sus sinónimos, y de nada sirve traducir a synsets las palabras monosémicas de la colección si no podemos asegurar una precisión aceptable en las palabras de la consulta, y viceversa.

Sin embargo, de los 91.272 synsets de WordNet 1.5 hay 43.128 en los que todos sus términos son monosémicos, es decir, sus términos sólo aparecen en ese synset. Estos synsets se denominarán aquí *synsets de variantes monosémicas* (SVM).

Hay 24.739 synsets de estas características que tienen más de un término, es decir, son synsets cuyos términos merecen ser representados en una colección mediante su índice de synset, pues no hay pérdida de precisión ni introducción de ruido en la desambiguación. Por otro lado, la traducción de estos términos a su SVM permite la convergencia de sinónimos sin pérdida de información.

Respecto a nombres, hay aproximadamente 60.500 synsets de los cuales 35.044 tienen todos sus lemas monosémicos. De ellos, hay 19.240 synsets que tienen más de un lema.

Además de aprovechar de forma inmediata las relaciones de sinonimia, a priori la expansión de SVM mediante synsets hiperónimos también resulta beneficiosa. Debido a la falta de ambigüedad del SVM los hiperónimos están perfectamente determinados.

3.5.2 Estadísticas en la colección de prueba *ohsumed*

La colección OHSUMED es una colección en el dominio médico que tiene unos 380Mb divididos en cinco ficheros correspondientes a un año cada uno entre 1987 y 1991. Sobre la base de 664.942 palabras pertenecientes al primero de ellos (ohsumed.87) se han obtenido las siguientes estadísticas relativas a SVM:

- 80.873 términos pertenecen a synsets SVM, lo que supone un 12'64%
- 36.829 términos pertenecen a synsets SVM con más de una variante, lo que supone el 5'54%.

Para obtener estas estadísticas se han tenido que detectar los términos multipalabra de WordNet contenidos en la colección. El número de términos multipalabra detectado en el fichero de la colección fue de 21.881.

Respecto al fichero de consultas, sobre una base de 1.516 palabras se han encontrado:

- 159 términos pertenecientes a synsets SVM, lo que supone un 10'49%.
- 67 términos pertenecientes a synsets SVM con más de una variante, lo que supone un 4'42%.

Sin embargo, de cara a la Recuperación de Información, los datos que interesan son los referentes a coincidencias entre consultas y documentos:

- 33 synsets SVM aparecen tanto en las consultas como en los documentos, implicando a un total de 5.575 ocurrencias de términos en la colección.
- De ellos, 16 synsets contienen palabras que no coinciden siempre en consultas y documentos, implicando a un total de 2971 ocurrencias de palabras en la colección de documentos. Por ejemplo,
    - En la consulta aparece *"ekg"* mientras que en los documentos aparece indistintamente "*ekg"* o "*electrocardiogram"*.
    - En la consulta aparece *"mri"*, mientras que en los documentos aparece tanto "mri" como la multipalabra "magnetic resonance imaging".

Es importante reseñar que palabras homógrafas de categorías distintas pueden introducir errores. Así, por ejemplo, *"who"* como nombre comparte synset SVM con "*world health organization*". Sin embargo, la inmensa mayoría de las ocurrencias de *"who"* en los documentos (677) no corresponde al SVM, sino al pronombre.

3.5.3 Conclusiones

Los *synsets de variantes monosémicas* (SVM) son un ejemplo de la utilidad de una indexación conceptual. Sustituyendo las ocurrencias de los términos que pertenecen a synsets SVM por el índice de synset es posible relacionar palabras sinónimas sin pérdida de precisión y con un coste de procesamiento despreciable.

Lamentablemente, las estadísticas obtenidas sobre la colección *ohsumed* muestran que el incremento de relaciones entre consultas y documentos que produce la consideración de SVM, no puede mejorar la recuperación de forma significativa.

Para abordar una indexación conceptual basada en synsets de WordNet será necesario tratar el problema de desambiguación del sentido de las palabras.

3.6 Recuperación multilingüe basada en indexación conceptual

La disponibilidad de las bases de datos léxicas de EuroWordNet (EWN) y de su versión en español, catalán y euskera (*ITEM*[7]) permite explorar una alternativa atractiva a la de traducción gracias al Índice InterLingua (ILI): la utilización de los *registros del índice InterLingua* (ILI-records) para indexar tanto las consultas como los documentos, y comparar conceptos en lugar de palabras clave. La *Figura 3-3* muestra la estructura de EWN y la interconexión de WordNets en diferentes idiomas a través del Índice InterLingua.

Las ventajas más apreciables de esta aproximación frente a la traducción a otro idioma son:

- La comparación entre documentos y consultas se hace a un nivel conceptual, evitando los problemas de polisemia de las palabras como términos de indexación, y permitiendo identificar términos sinónimos como el mismo término de indexación.

- La comparación entre consultas y documentos se realiza en un espacio independiente del idioma, simplificando el problema de fusionar resultados de varias búsquedas monolingües en varias colecciones textuales distintas. Todos

[7] Proyecto ITEM: Recuperación de Información Textual en un Entorno Multilingüe (CICyT TIC96-1243-C03-01)

los textos pueden ser indexados con los mismos términos de indexación, sin importar el idioma original en que han sido escritos.

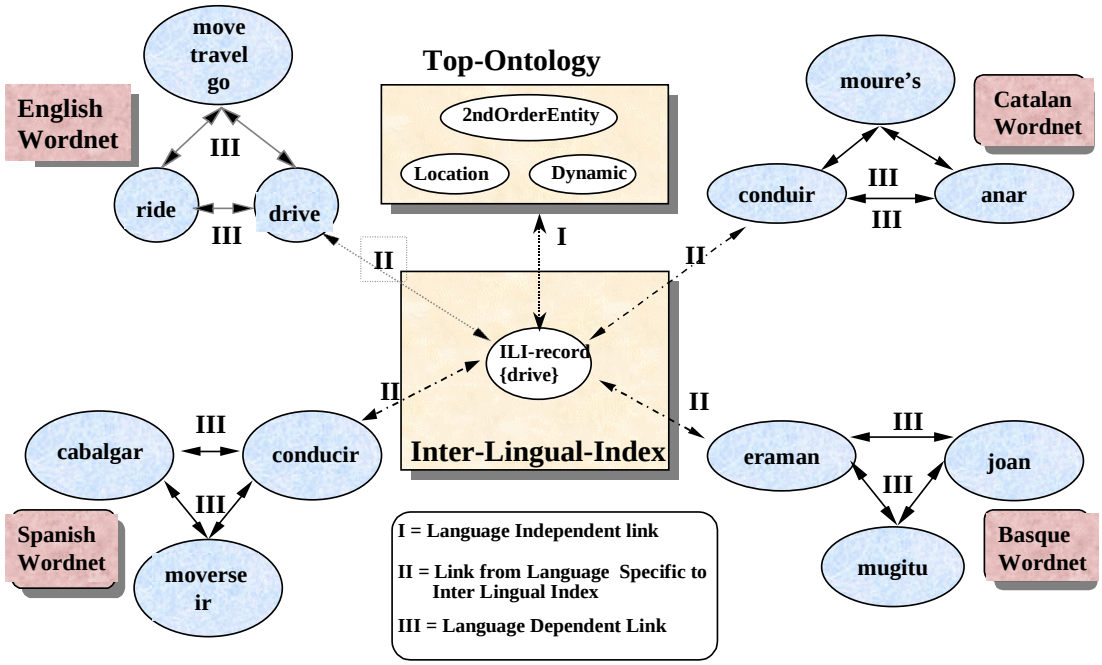


Figura 3-3. Estructura de EuroWordNet

3.7 Viabilidad de una recuperación basada en indexación conceptual

El siguiente experimento aborda la cuestión de cuál es el grado de error permisible en desambiguación del sentido de las palabras (WSD) para que el modelo de recuperación basado en la indexación de synsets suponga una mejora en la recuperación.

A partir de la colección IR-SEMCOR y con el fin de obtener unos resultados comparables con los de (Sanderson 1994), se ha preparado una segunda colección en la que tanto consultas como documentos en texto, ambos sin etiquetas, son

transformados a pseudo-palabras de tamaño 5. Se ha elegido un tamaño 5 de las pseudo-palabras (al igual que en los experimentos de Sanderson) porque la polisemia media de los términos en SEMCOR es también 5.

3.7.1 Sensibilidad a los errores de desambiguación

En uno de sus experimentos, (Sanderson 1994) desambigua la colección de pseudo-palabras de tamaño 5 introduciendo tasas fijas de error. Así, la colección original (con palabras en vez de pseudo-palabras) corresponde a una desambiguación del 100% de precisión. En su colección, y con pseudo-palabras, Sanderson llegó a la conclusión de que una precisión en la desambiguación por debajo del 90% producía peores resultados en la recuperación que no desambiguar. Su conclusión fue que la WSD necesitaba ser extremadamente precisa para mejorar la recuperación.

3.7.2 Definición del experimento

El siguiente experimento tiene como objetivo determinar la tasa de errores permisible en WSD para que la indexación mediante synsets de WordNet mejore la recuperación o, al menos, no la empeore y permita así una recuperación multilingüe.

En primer lugar, se introducirán tasas fijas de error de desambiguación sobre pseudo-palabras en la colección construida a tal efecto sobre IR-SEMCOR.

En segundo lugar, puesto que disponemos de una colección con los sentidos de las palabras desambiguados manualmente, también vamos a introducir tasas fijas de error en esta colección, y así evaluar la sensibilidad de la recuperación ante una desambiguación real. Esto permitirá comparar el comportamiento real de la desambiguación respecto a la desambiguación de pseudo-palabras.

En este experimento, se define tasa de error como el porcentaje de palabras polisémicas desambiguadas incorrectamente.

3.7.3 Realización del experimento y resultados

Los resultados del experimento pueden observarse en la *Figura 3-4* en la que el eje de ordenadas corresponde a la precisión media de la recuperación medida en 10 puntos de *recall*, frente al porcentaje creciente de errores en el eje de abscisas.

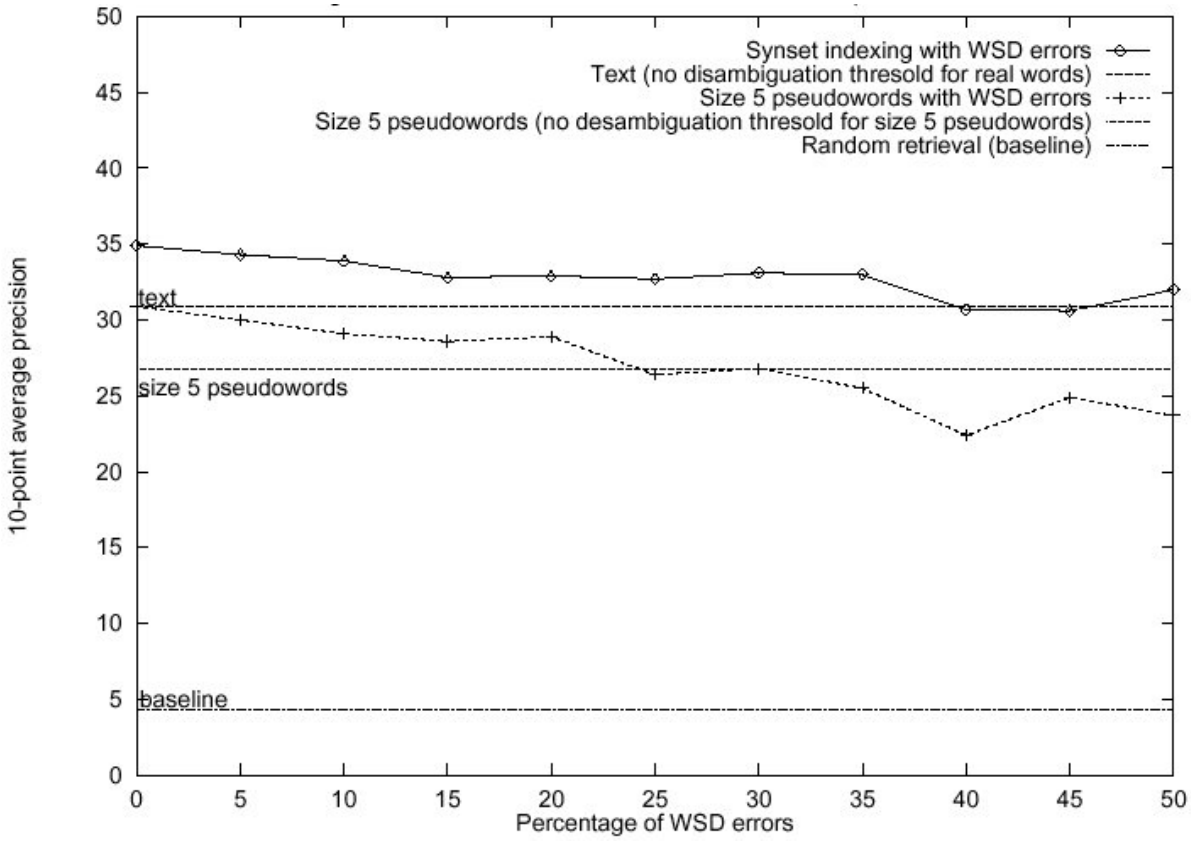


Figura 3-4. Pérdida de precisión frente al porcentaje de errores en WSD

Por una parte, la gráfica compara la precisión de la recuperación con errores de desambiguación de pseudo-palabras de tamaño 5 frente a la precisión de la recuperación manteniendo las pseudo-palabras sin desambiguar. Puede observarse que IR-SEMCOR es una colección más resistente a los errores de desambiguación que la colección de Reuters utilizada por Sanderson. El umbral del 90% se ve reducido al 75% de precisión, es decir, con una tasa de error del 25% o menos en la desambiguación de pseudo-palabras, la recuperación mejora.

Por otra parte, la gráfica también compara la precisión de la recuperación con errores de WSD real frente a la recuperación con textos sin desambiguar. Puede observarse que la indexación mediante synsets es mucho más tolerante a los errores de WSD. A partir de una precisión del 60% en la desambiguación, es posible mejorar la recuperación respecto a texto sin desambiguar.

De la gráfica se desprende además una discrepancia entre el comportamiento de pseudo-palabras y de palabras respecto a la recuperación. La razón de esta discrepancia debe buscarse en:

1. Mientras que las componentes de una pseudo-palabra son palabras con significados totalmente diferentes, en realidad las palabras polisémicas tienen sentidos que están relacionados entre sí. En (Buitelaar 1998) se estima que sólo un 5% de entradas para una palabra en WordNet pueden considerarse verdaderos homónimos, es decir, sentidos no relacionados, mientras que el 95% de los sentidos pueden considerarse una extensión de un sentido nuclear. Así pues, un error de desambiguación puede ser menos perjudicial si se ha elegido un sentido relacionado con el sentido correcto. Este hecho sugiere que para tareas de Recuperación de Información no es necesaria un desambiguación completa, sino que posiblemente una ponderación de los sentidos más probables sea suficiente siendo únicamente necesario descartar los sentidos menos probables. Esta aproximación subyace a (Schütze 1995) donde se mejora en un 14% la recuperación desambiguando las palabras con un tesauro construido sobre la base de co-ocurrencias entre palabras.

2. La indexación mediante synsets no sólo desambigua palabras, sino que permite recuperar documentos que contienen palabras sinónimas. Esto produce una mejora de la cobertura (*recall*) sin dañar la precisión y, por tanto, mejora la recuperación. Es decir, la desambiguación de pseudo-palabras tiene un comportamiento similar a la desambiguación del sentido de las palabras más que a la utilización de synsets.

3.7.4 Conclusiones

El hecho de que la recuperación basada en indexación con synsets sea resistente a un 40% de error en el proceso de desambiguación abre la posibilidad de que la desambiguación se realice de forma automática pues se trata de un grado de precisión no demasiado alejado del estado de la cuestión en WSD.

Pero al margen de una mejora en la recuperación que debido a la tasa de errores de desambiguación automática será una mejora reducida, el empleo de synsets como índices de recuperación abre la posibilidad de realizar una recuperación multilingüe sin la pérdida de precisión que se produce al traducir consultas o documentos al idioma destino.

Por estas razones, y tras los estudios expuestos de viabilidad del modelo, se decidió implementar el modelo de indexación conceptual dentro del buscador multilingüe del proyecto ITEM (ITEM Search Engine). Este buscador se describe más adelante.

3.8 El motor de búsqueda ITEM

La evaluación cualitativa de la aproximación de recuperación multilingüe basada en indexación conceptual, así como otras alternativas interesantes que proporcionaba el uso de EuroWordNet, fue realizada sobre el motor de búsqueda ITEM implementado a tal efecto por el grupo PLN de la UNED en un trabajo colectivo.

El motor de búsqueda ITEM es una contribución que permite realizar tests cuantitativos y cualitativos sobre el impacto de las bases de datos léxicas y las herramientas de procesamiento de lenguaje natural en los sistemas de recuperación de información mono y multilingües y, en particular, las distintas estrategias de desambiguación semántica (WSD) para la indexación conceptual y la expansión de la consulta en otros idiomas.

Este motor implementa dos enfoques alternativos para la recuperación de información multilingüe (en español, inglés y catalán). El primero consiste en traducir la consulta de su idioma original a los otros dos idiomas de búsqueda posibles a través del Índice InterLingua de EuroWordNet, y en realizar a continuación tres procesos de búsqueda monolingüe con el motor de búsqueda standard INQUERY (Callan 1992). El segundo, como ya se ha mencionado, se basa en indexación conceptual sobre synsets.

3.8.1 Traducción de la consulta mediante EuroWordNet

La traducción de la consulta mediante EuroWordNet es un enfoque que se aproxima a la recuperación multilingüe basada en diccionarios, donde cada palabra original se sustituye por las traducciones obtenidas en un diccionario bilingüe, después de ciertos filtros estadísticos (en especial para traducir expresiones multipalabra). Sin embargo, el uso de la red semántica EWN/ITEM ofrece cierto número de ventajas.

1. Los wordnets de español, catalán e inglés juegan el papel de seis diccionarios bilingües. La ventaja de disponer de un índice InterLingua crece rápidamente con el número de idiomas contemplados, y el número potencial de idiomas para el motor de búsqueda es actualmente 10 (inglés, español, catalán, euskera y el resto de idiomas de EWN: holandés, italiano, francés, alemán, estonio y checo).

2. La desambiguación semántica se puede realizar explícitamente en un nivel independiente del idioma (la representación del índice InterLingua). La desambiguación proporciona los registros del ILI adecuados, y los registros

del ILI están ligados a los conjuntos de palabras sinónimas en cada idioma contemplado.

3. Las relaciones semánticas en la base de datos léxica EWN/ITEM permiten una expansión controlada con términos semánticamente relacionados: hipónimos, merónimos, etc. Las relaciones de hiperonimia/hiponimia en el índice InterLingua permiten obtener traducciones aproximadas para los términos de la consulta que no tienen equivalentes en el (o los) idiomas objetivo. Por ejemplo, ``governor's race'' no tiene equivalente en español. Sin embargo, ``governor's race'' se puede ligar a ``elecciones'' a través de ``elections'', que es el hiperónimo directo de ``governor's race''. Otro ejemplo es ``grand jury'', que no tiene equivalente en español pero puede tener como traducción aproximada ``jurado'', como un equivalente en español para el concepto ``jury'', que es hiperónimo directo de ``grand jury''.

3.8.2 Indexación conceptual

La aproximación basada en indexación conceptual requiere que tanto los documentos como las consultas sean procesados por una cascada de analizadores léxicos, en la que los lematizadores y etiquetadores de categoría gramatical son dependientes del idioma, pero el resto de procesos son independientes de la lengua.

En la aproximación por *traducción de la consulta* se realiza el mismo tipo de procesamiento pero únicamente sobre las consultas. La secuencia de procesos es la siguiente:

3.8.2.1 Lematización y etiquetado de categorías gramaticales

El español y el catalán son procesados con el analizador morfológico MACO+ y el etiquetador RELAX (Márquez 1997;Carmona 1998). El inglés se procesa con una versión del Brill tagger (Brill 1992) y con el lematizador de WordNet 1.5 (Miller 1990).

3.8.2.2 Detección de expresiones multipalabra

La detección de expresiones multipalabra es, en este enfoque, una tarea independiente del idioma que considera sólo las expresiones incluidas en la base de datos léxica EWN/ITEM.

3.8.2.3 Desambiguación del sentido de las palabras

El motor de búsqueda ofrece tres opciones para desambiguar las palabras (tanto en documentos como en consultas):

1. **Primer sentido**: se toma el primer sentido de la base de datos léxica EWN/ITEM. En el WordNet inglés, el primer sentido corresponde con el sentido más frecuente en Semcor. La elección del sentido más frecuente de una palabra es la mejor aproximación si no se realiza un procesamiento adicional. Sin embargo, no siempre se disponen de estadísticas para decidir cuál es el sentido más frecuente. Esto es lo que ocurre con el resto de idiomas del buscador ITEM. En los wordnets español y catalán el primer sentido no tiene por qué ser necesariamente el más frecuente.

2. **Todos los sentidos**: se toman todos los posibles sentidos de cada nombre en los documentos como elementos de indexación igualmente válidos.

3. **Densidad conceptual**: en esta opción, todos los nombres en el documento (o en la consulta) se desambiguan, asignando probabilidades diferentes para cada sentido. La desambiguación se realiza mediante una implementación eficiente de un algoritmo no supervisado inspirado en (Agirre 1996) que utiliza únicamente información jerárquica y medidas de distancia conceptual para realizar la desambiguación (Fernández-Amorós 2001).

3.8.2.4 Representación mediante synsets

Una vez se determinado el sentido de una palabra, ésta se sustituye por el índice numérico del synset al que pertenece dicho sentido en EuroWordNet. Esto permite que una consulta recupere documentos que contienen términos sinónimos.

3.8.3 Interfaz del buscador multilingüe ITEM

El motor de búsqueda ITEM proporciona una experiencia directa con el uso de una red semántica multilingüe, así como con las peculiaridades de usar un procesamiento léxico exhaustivo. La interfaz web al motor de búsqueda permite al usuario ajustar parámetros relacionados con el procesamiento de lenguaje natural de la consulta (y los documentos), refinar los resultados del procesamiento de la consulta y comparar resultados con distintos parámetros.

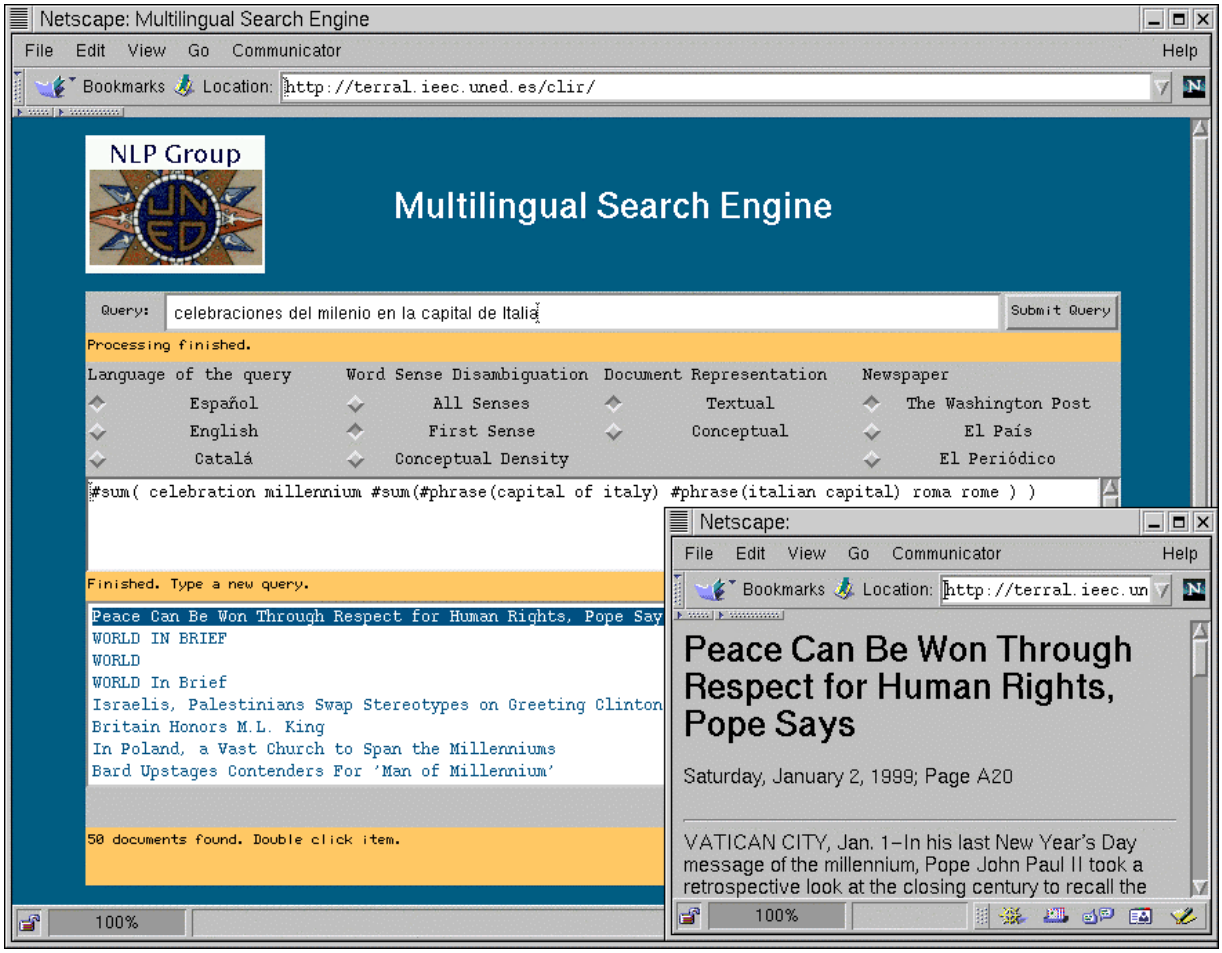


Figura 3-5 Interfaz del motor de búsqueda ITEM

En la *Figura 3-5* se puede ver el aspecto de la interfaz web. La caja de texto superior de la interfaz se usa para realizar la consulta. Las opciones de procesamiento se seleccionan en los botones que están inmediatamente debajo. A continuación se describe el efecto de cada una de ellas.

1. **Idioma de la consulta**. Las posibilidades son: inglés, español o catalán.

2. **Idioma de los documentos**. Depende del periódico que se selecciona: español para "El País", inglés para el "Washington Post" y catalán para "El Periódico".

3. **Representación de los documentos**. Las opciones son "textual" o "conceptual". En la representación textual los textos se mantienen en texto, mientras que en la representación conceptual, los documentos se han procesado lingüísticamente de la forma mencionada anteriormente hasta su desambiguación semántica y su traducción a synsets. En la opción textual,

la consulta se traduce al idioma de la colección a través del índice InterLingua (a modo de diccionario bilingüe), y entonces se realiza una búsqueda estándar sobre la base documental original en formato texto. En la opción conceptual, el procesado de la consulta se detiene al nivel de representación conceptual, y se compara con la representación conceptual de documentos en términos del índice InterLingua.

4. **Desambiguación conceptual**. El usuario puede elegir entre considerar todos los posibles sentidos de cada palabra (All Senses), tomar siempre el primer sentido (First Sense) o usar el algoritmo de desambiguación mencionado más arriba basado en densidad conceptual (Conceptual Density). Si la representación del documento seleccionada es "textual", el criterio de desambiguación seleccionado sólo afecta a la traducción de la consulta, restringiendo la traducción de la misma a los conceptos determinados por el método de desambiguación. Si el método de desambiguación es el "primer sentido" o "todos los sentidos", entonces la consulta se procesa hasta la desambiguación correspondiente y la búsqueda se realiza sobre la colección procesada de igual modo. Si el método de desambiguación elegido es "densidad conceptual", entonces el procesamiento de la consulta difiere un poco del procesamiento de los documentos. Para los documentos, se toma el sentido al que se le ha asignado mayor probabilidad, y todos aquellos que tengan al menos el 80% de su valor, mientras que el resto se descartan. Para las consultas, en las que el contexto es, usualmente, demasiado pequeño para una desambiguación fiable, se conservan todos los sentidos pero pesándolos de acuerdo con sus probabilidades asignadas.

5. **Detección de expresiones multipalabra**. Cuando se activa, las expresiones multipalabra en documentos y consultas se toman como unidades de indexación únicas. Un sintagma se considerará unidad de indexación sólo en el caso de compuestos exocéntricos, como "fisher cat", en los que el significado de los componentes no está relacionado con el significado de la expresión completa. Otras expresiones, como "abstract art", se tratan combinando los significados de las palabras que las componen.

Una vez que se procesa la consulta, el sistema proporciona los siguientes resultados:

1. En el área de texto bajo los botones de selección, se muestra la consulta expandida de acuerdo con las opciones de búsqueda y el lenguaje del motor de búsqueda utilizado (INQUERY). El usuario puede refinar esta consulta expandida añadiendo y eliminando términos (o conceptos) en esta caja y pedir una nueva búsqueda ésta vez sin más procesamiento léxico.

2. En el área inferior se muestra una lista ordenada de documentos relevantes para la consulta en el periódico seleccionado. El usuario puede pulsar en el título para ver el texto completo.

3.8.4 Ejemplo de funcionamiento del buscador multilingüe ITEM

Para observar el funcionamiento del sistema tómese como ejemplo la siguiente consulta en español,

celebraciones del milenio en la capital de Italia

Sobre la consulta se realiza el procesamiento léxico ya descrito cuyo resultado se muestra en la *Figura 3-6.*

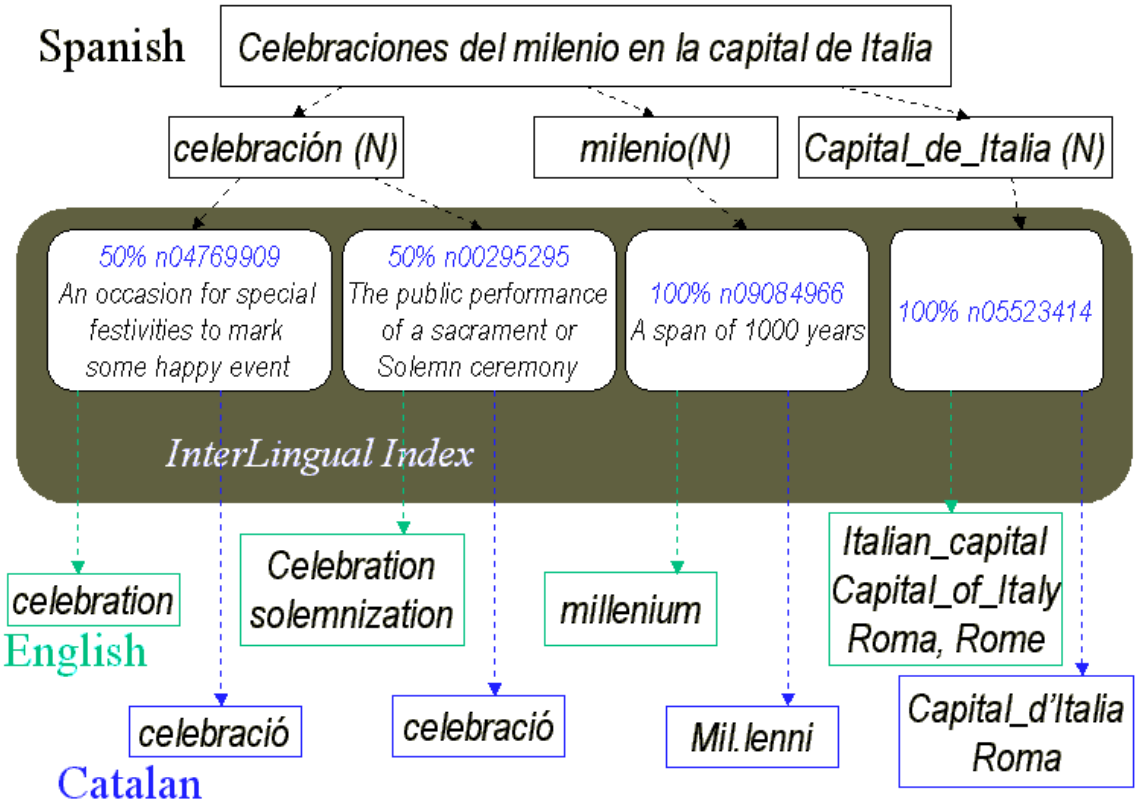


Figura 3-6. Ejemplo de procesamiento léxico de una consulta.

Los pasos principales de procesamiento son:

1. Identificación de expresiones multipalabra, lemas y categoría gramatical adecuada.
2. Representación en términos del índice InterLingua, como probabilidades asignadas al algoritmo de desambiguación
3. Expansión en los idiomas objetivo.

La información obtenida en los pasos 2 y 3 se usa para construir la consulta final de acuerdo con las opciones seleccionadas por el usuario. Por ejemplo, cuando la representación documental es ``textual'' y la opción de desambiguación es *"todos los sentidos",* el resultado es:

#sum(#sum(celebration #sum(celebration solemnization)) millennium
#sum(#phrase(capital of italy) #phrase(italian capital) roma rome))

Cuando la opción de desambiguación es *"primer sentido"*, el resultado es:

#sum(celebration millennium #sum(#phrase(capital of italy)
#phrase(italian capital) roma rome))

Si la opción de desambiguación es *"densidad conceptual",* se usan los pesos en la construcción de la consulta:

#sum(#wsum(100 50 celebration 50 #sum(celebration solemnization).
#wsum(100 100 millennium)
#wsum(100 100 #sum(#phrase(capital of italy) #phrase(italian capital)
roma rome)))

Si la representación de los documentos es *"conceptual"*, la consulta se representa mediante synsets. Por ejemplo, si la estrategia de desambiguación fuera "*primer sentido"*, la consulta se convertiría en:

#sum(n04769909 n09084966 n05523414)

donde, por ejemplo, n05523414 representa el registro del índice InterLingua:

n05523414
English: Rome, Roma, Italian capital, capital of Italy
Spanish: capital de Italia, Roma
Catalan: capital d'Itàlia, Roma

=> hypernym: n05483778
English: national capital
Spanish: capital de nación
Catalan: capital de nació

Tras este procesamiento, el usuario puede refinar la consulta, ya sea reformulando la consulta original o, lo que es más interesante, añadiendo o eliminando términos directamente de la consulta expandida. Por ejemplo, el usuario puede escoger una expansión "todos los sentidos", después eliminar manualmente aquellas traducciones que no son apropiadas, y consultar directamente a la base textual con el resultado.

3.8.5 Evaluación cualitativa

La experiencia directa utilizando la interfaz de búsqueda permite extraer algunas conclusiones sobre la calidad de los recursos y herramientas empleados, y sobre la utilidad de estos recursos en recuperación de información multilingüe.

Para traducción/expansión de consultas, las bases de datos de EuroWordNet y EWN/ITEM ofrecen características interesantes por comparación con los diccionarios bilingües. Las relaciones semánticas en el Índice Interlingua permite encontrar traducciones aproximadas cuando no se puede encontrar una equivalencia directa (o no existe en el idioma objetivo), y permite sugerir otros términos relacionados semánticamente. Sin embargo, como en el caso de los diccionarios electrónicos, es necesario adaptar el sistema al dominio para obtener traducciones adecuadas a términos y significados propios del dominio, especialmente con expresiones multipalabra.

La indexación conceptual es una opción atractiva, a priori, para realizar recuperación multilingüe. Pero deben resolverse cuatro retos principales:

1. Los sentidos para una palabra dada considerados en la base de datos léxica, deberían reflejar diferencias de uso en contexto. De no ser así, excesivas distinciones en los sentidos sólo perturban negativamente el proceso de recuperación de información, añadiendo excesivo ruido. Debido a esta excesiva granularidad de EWN/ITEM es necesario encontrar formas de agrupar los sentidos de EWN/ITEM para los propósitos de Recuperación de Información (Chugur 2000;Gonzalo 2000).

2. EWN/ITEM es una base de datos léxica construido con propósito general y al margen de un dominio concreto de aplicación. Esto provoca la consideración

de sentidos que no existen en el dominio de la colección, por una parte, y por otra, la falta de conceptos específicos, propios del dominio. Para superar este problema, es necesario desarrollar métodos de adaptación de la red semántica a una colección concreta, así como enriquecer la red con etiquetas de dominio.

3. La desambiguación semántica es todavía un tema de investigación abierto especialmente cuando la tarea a realizar es una anotación semántica exhaustiva de los nombres y verbos en una colección de textos en tres idiomas distintos. El algoritmo basado en densidad conceptual satisface los requisitos de cobertura, ya que es no supervisado e independiente del idioma (aunque sí requiere que el idioma tenga una base de datos léxica tipo WordNet). Sin embargo, la interacción con el sistema ha demostrado que no es suficientemente preciso.

4. Las unidades de indexación que proporciona la red semántica son unidades léxicas demasiado pequeñas para la traducción y, por ello, en la indexación se produce una pérdida de información.

3.9 Conclusiones

Los resultados obtenidos en los experimentos de estudio de viabilidad llevaron a la decisión de implementar un prototipo que permitiera evaluar no sólo cuantitativamente, sino también cualitativamente una recuperación multilingüe basada en indexación conceptual sobre synsets de EuroWordNet. La evaluación de este sistema arroja a la luz tres grandes retos que de momento no hacen efectiva la recuperación basada en indexación conceptual: la excesiva distinción de sentidos de EWN, la adaptación de EWN a dominios específicos, y el desarrollo de métodos más precisos de desambiguación del sentido de las palabras. Además, las unidades léxicas de EWN no parecen ser unidades de traducción apropiadas.

A esto hay que añadir que el coste computacional que acarrea la cascada de procesamiento lingüístico tanto de los documentos como de la consulta cuestiona su uso en colecciones medianas (100.000 documentos).

Por otro lado, los experimentos que tratan de incorporar en el modelo tradicional de recuperación de documentos técnicas lingüísticas como el etiquetado de categoría gramatical o la indexación de sintagmas tampoco resultan satisfactorios.

De todo esto se concluye que, de momento, el papel de NLP en IR pasa por superar el modelo de acceso a la información como recuperación y ordenación de una lista de documentos. Es necesario encontrar nuevos paradigmas con capacidad para integrar, por una parte, la información que se pueda extraer de forma automática mediante técnicas lingüísticas pero, por otra, la información que proporcionan los usuarios mediante sus consultas, sus elecciones, y su interacción con el sistema. Es decir, la subordinación de las técnicas lingüísticas a la tarea de recuperación de documentos no parece aportar mejoras significativas en el acceso a la información. Será en la redefinición apropiada de las tareas más próximas al nivel del usuario donde las técnicas NLP puedan ayudar a mejorar el acceso a la información.

Capítulo 4

Acceso interactivo a la información mediante sintagmas

Todo el trabajo expuesto en el capítulo anterior confirma la tesis de algunos autores (Sparck Jones 1999) de que las técnicas lingüísticas aplicadas a la indexación en la tarea de *recuperación de documentos* no parecen aportar mejoras significativas en el acceso a la información. Por tanto, debe plantearse el papel que pueden desempeñar las técnicas lingüísticas en tareas más próximas al nivel del usuario. Superar el modelo de *recuperación de documentos* requiere un replanteamiento del concepto de *acceso a la información*.

La redefinición de las tareas de acceso a la información en las que el Procesamiento del Lenguaje Natural puede realizar alguna aportación significativa, pasa por el replanteamiento de la siguiente pregunta: *¿Cómo ayudar al usuario a expresar, precisar y satisfacer sus necesidades de información?*

En este trabajo se replantea la pregunta en términos lingüísticos, *¿cómo ayudar al usuario a contextualizar su consulta?*

Y en términos cognitivos, *¿qué inferencias puede y debe realizar el sistema y cuáles el usuario en la interpretación de la consulta?*

La hipótesis que se asume en este trabajo es que:

> *la mejor interpretación de una consulta es el conjunto de fragmentos que componen la información que se busca.*

Por tanto, el proceso de definición, concreción y contextualización de una consulta, es decir, su interpretación, se identifica con el propio proceso de acceso a la información.

En este trabajo se asume que el único que puede otorgar significado a la información es el usuario, y que el sistema sólo puede ofrecerle un procesamiento parcial de la información para ayudarle a recorrer el camino de forma más eficiente. Todo esto conduce a la necesidad de considerar la interactividad en los modelos de acceso a la información y conjugarla con un procesamiento inferencial automático.

En este capítulo se propone un modelo interactivo de acceso a la información que incorpora inferencias lingüísticas sobre sintagmas (4.1) para tratar los problemas de variación terminológica y multilingüismo en el acceso a la información.

La arquitectura que se propone se basa en la extracción y uso de sintagmas tanto para enriquecer y traducir las consultas como para acceder a la información desde los sintagmas sugeridos por el sistema. Esta arquitectura afecta al modelo de indexación, al de recuperación y al modelo de interacción. En primer lugar, el proceso de indexación (4.2) de una colección se dirigirá a construir de forma automática los vocabularios de la colección contemplando no sólo palabras aisladas sino también sintagmas, independientemente de los idiomas de la colección. En segundo lugar, el proceso de recuperación (4.3) se realizará sobre la base del vocabulario extraído, sirviendo como referencia para aceptar o desechar variaciones morfosintácticas, semánticas y translingües de la consulta. En tercer lugar, los sintagmas así obtenidos y recuperados a partir de la consulta se organizarán en una nueva área que ofrece al usuario un mínimo de interacción (4.4) dirigida a seleccionar los subconjuntos del vocabulario que se ajustan más a sus necesidades de información, proporcionándole acceso directo a los documentos que lo contienen. A continuación se describe con detalle este modelo.

4.1 Inferencia sobre sintagmas

Muchas de las consultas que se hacen a un buscador se pueden entender como la búsqueda de un concepto o una entidad. Aunque más o menos compleja, la expresión lingüística de un concepto se corresponde con el sintagma[8] nominal.

[8] Con sintagma nos referimos a un grupo de palabras que expresan un concepto subordinado a la oración. Definido de esta manera, la idea de sintagma es más general que la de *compuesto léxico* o la de *colocación* que hacen explícita la necesidad de que el grupo de

La consideración de sintagmas impone una restricción de adyacencia sobre las palabras de la consulta que afecta negativamente al índice de recuperación (recall). Sin embargo, (Pickens 2000) muestran que cuanto mayor sea el número de documentos recuperados según el modelo booleano (i.e. sin imponer la restricción de adyacencia), mayor es el valor del sintagma para identificar los documentos verdaderamente relevantes. Esto significa que cuanto mayor sea la colección más útil se hace la consideración de sintagmas. Este es precisamente el caso en el que se hace necesario construir niveles intermedios de acceso a la información. En este trabajo, este nivel está constituido precisamente por los sintagmas extraídos automáticamente de la colección y que están directamente relacionados con la consulta.

El índice de recuperación (*recall*) se eleva mediante la expansión de la consulta, pero con el perjuicio de una disminución en la *precisión* de la recuperación. Sin embargo, la consideración de sintagmas permite reducir la ambigüedad de la expansión (Ballesteros 1998) y mantener niveles altos de precisión. La simple co-ocurrencia de las palabras de expansión y/o traducción en un mismo sintagma se convierte en un criterio muy potente para desechar combinaciones carentes de sentido. Esta aproximación va a permitir abordar los problemas de variación morfosintáctica, semántica y translingüe.

Finalmente, una de las ventajas más importantes de considerar sintagmas es que gracias a su lectura y comprensión como conceptos precisos y poco ambiguos, el usuario puede identificar rápidamente información relevante. El sintagma es la unidad más manejable y a la vez significativa que se puede utilizar para proporcionar información sobre los documentos sin necesidad de explorarlos. Es decir, el sintagma adquiere un papel relevante cuando se trata de acceso interactivo a la información.

Existen modelos de recuperación que consideran la proximidad de las palabras sin exigir la adyacencia de las mismos y, de esta manera, aumentar la precisión sin disminuir el índice de recuperación. Sin embargo, el enfoque de este trabajo no es mejorar el ranking de documentos sino proporcionar al usuario un nivel intermedio

---

palabras se haya lexicalizado o, al menos, co-ocurra frecuentemente. Sin embargo, con fines de acceso a la información, el grado de lexicalización de un sintagma será uno de los criterios más importantes para determinar la relevancia de un sintagma. Una expresión lexicalizada y con mayor peso conceptual tiene mayor capacidad descriptiva y discriminativa de la información que se busca. Por ello, en lo referente a extracción de terminología, también nos referiremos a los sintagmas como términos poli-léxicos o simplemente términos.

de acceso a la información abriendo la posibilidad de que una interacción con el sistema le lleve de forma efectiva a los documentos que busca.

4.2 Modelo propuesto de indexación

El modelo propuesto de indexación afecta a dos subtareas: la extracción de los términos (palabras y sintagmas) y la construcción de los índices que den acceso eficiente tanto a los sintagmas como a los documentos.

La consideración de sintagmas en la indexación de grandes colecciones de documentos acarrea serios problemas de eficiencia en dos aspectos: el tiempo de procesamiento que se requiere para la extracción de sintagmas, y el tamaño de los índices cuando se consideran sintagmas. Respecto a la recuperación de la terminología relacionada con la consulta es necesario considerar el tiempo de obtener, seleccionar y organizar los términos (palabras y sintagmas) relacionados con la consulta. Estos problemas de eficiencia son los que cuestionan, en muchas ocasiones, el uso de técnicas lingüísticas en Recuperación de Información. Por ello, en los apartados siguientes se prestará atención a estos aspectos, comparando las distintas alternativas que permitan alcanzar un equilibrio entre eficiencia y resultados.

4.2.1 Indexación de sintagmas en IR

Si los índices de recuperación contienen las posiciones que ocupan las palabras en los textos, entonces la recuperación de los documentos que contienen un sintagma es relativamente sencilla. Consideremos el ejemplo de índice ya comentado en el apartado 2.1.1.2:

Número	Término	(Documento; posiciones)
1	universidad	(1;2,14) (4;8)
2	distancia	(1;6) (3;9)
3	lenguaje	(2;3)
4	natural	(2;4) (9;5,16, 27)
…	…	…

Una consulta con el sintagma “lenguaje natural” recuperaría el documento 2 puesto que *lenguaje* y *natural* aparecen en el documento 2 en posiciones adyacentes. De igual forma, se puede plantear la tarea de recuperación de terminología. Por

ejemplo, ¿existe en la colección algún sintagma que contenga *lenguaje* y *natural*? Evidentemente sí, puesto que hay un documento que contiene *lenguaje* y *natural* en posiciones adyacentes. ¿Hay más sintagmas que contengan estas palabras? No, puesto que no coinciden en más documentos.

Sin embargo, analicemos la siguiente tarea, ¿existe en la colección algún sintagma que contenga *universidad* y *distancia*? Según el índice ambas palabras aparecen en el documento 1:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
*	universidad	*	*	*	distancia	*	*	*	*	*	*	*	universidad	*	*

Para responder a la pregunta sería necesario explorar el documento, a menos que se disponga previamente de la información. Por ejemplo, supongamos que el texto ha sido segmentado antes de indexarse de forma que se puede considerar que si dos palabras aparecen en el mismo segmento entonces hay un sintagma nominal que los contiene:

Número	Término	(Documento; segmento; posiciones)
1	universidad	(1; 2; 2) (1; 4; 14) (4; 3; 8)
2	distancia	(1; 2; 6) (3; 3; 9)
3	lenguaje	(2; 1; 3)
4	natural	(2; 1; 4) (9; 2; 5) (9; 4; 16) (9; 6; 27)
…	…	…

De esta forma, se podría afirmar que, efectivamente, hay un sintagma que contiene a *universidad* y *distancia*, pero si quisiéramos saber cuál para ofrecerlo al usuario como sugerencia tendríamos que explorar el documento.

La ventaja de esta aproximación es que los índices mantienen un tamaño razonablemente pequeño. La desventaja es que la recuperación de sintagmas relacionados con la consulta exige la lectura de muchos documentos con el consiguiente coste computacional. Por ejemplo, supongamos que *"universidad nacional de educación a distancia"* aparece en 200 documentos. La consulta *"universidad, distancia"* obligaría a leer los 200 documentos para sugerir un único sintagma puesto que no es posible establecer sin leer el documento cuáles son las palabras intermedias entre *universidad* y *distancia*. Si en vez de un sintagma se ofrecen decenas de ellos como sugerencias posibles, entonces el coste computacional se hace inviable.

La alternativa que se va a seguir en este trabajo es indexar los sintagmas. Al igual que un índice de documentos nos permite saber que documentos contienen un determinado término, un índice de sintagmas nos permitirá saber qué sintagmas contienen una determinada palabra. La ventaja de esta aproximación es que se reduce considerablemente el tiempo en recuperar los sintagmas relacionados con una consulta. La desventaja será que los índices pueden llegar a ser muy grandes y será necesario establecer criterios de selección de sintagmas.

4.2.2 Extracción de sintagmas

El número de sintagmas a considerar en los índices y el tiempo para extraerlos está relacionado con las técnicas de extracción así como su precisión. Las dos aproximaciones generales de extracción de sintagmas provienen de cada una de las comunidades que convergen en el área de acceso a la información mediante *exploración de sintagmas* (*Phrase Browsing*). Por una parte, la comunidad de Recuperación de Información (IR) explota los algoritmos de tratamiento de cadenas alfanuméricas considerando información estadística para determinar cuándo una cadena constituye un sintagma. Por otra parte, la comunidad de Procesamiento del Lenguaje Natural (NLP) explota el uso de herramientas y conocimiento lingüístico para identificarlos.

Aunque la aproximación NLP permite obtener sintagmas de mayor calidad, supone, en general, un coste computacional tan elevado que resulta imposible indexar colecciones de tamaño TREC (2 Gb). Por ello, desde la perspectiva de la comunidad de IR, es común definir un sintagma sencillamente como una secuencia de palabras que aparece más de una vez en la colección, y añadir ciertas restricciones que ayuden a garantizar la calidad de los sintagmas extraídos (el sintagma no contenga delimitadores como signos de puntuación, el sintagma no empiece ni termine con una stopword, etc.) (Paynter 2001b).

De esta forma, el problema de identificar sintagmas e inferir jerarquías de sub-sintagmas sin utilizar técnicas lingüísticas se ha abordado desde hace tiempo. (Ziv 1978) ya extraían de una cadena de entrada los subsintagmas y los organizaban de forma que era posible el acceso eficiente a los super-sintagmas de una palabra o sintagma en una dirección. RE-PAIR (Wolff 1975;Wolff 1980) y SEQUITUR (Nevill-Manning 1999) son sistemas que extraen jerarquías de subsintagmas permitiendo su exploración en ambas direcciones. Todos estos algoritmos segmentan la cadena de entrada en secuencias de palabras que no se solapan consiguiendo que el coste computacional sea lineal sobre la longitud de la cadena de entrada.

Sin embargo, la segmentación del texto de entrada no es una buena política para identificar jerarquías de subsintagmas. Por ejemplo, la segmentación del texto "context free grammar" puede realizarse de dos maneras (por la izquierda o por la derecha), dando lugar a jerarquías diferentes:

1. (context free) (grammar)

context
context free
context free grammar
...
context free language

2. (context) (free grammar)

grammar
free grammar
context free grammar
...
recursion free grammar

A partir de la primera segmentación, el sintagma "context free" nos permite acceder al super-sintagma "context free language" que resultaría inaccesible con la segunda segmentación. Sin embargo, la segunda segmentación nos permite acceder al super-sintagma "recursion free grammar" que no es accesible con la primera segmentación.

(Paynter 2001b) describe un algoritmo en varias fases para inferir jerarquías de subsintagmas en ambas direcciones (izquierda y derecha) considerando el solapamiento de sintagmas, y sin utilizar técnicas lingüísticas.

Por otro lado, un análisis sintáctico parcial basado en el ajuste de patrones sintácticos también permite la identificación de sintagmas considerando su solapamiento.

Una ventaja importante de la aproximación IR de cara a la recuperación multilingüe, es que el tratamiento de cadenas alfanuméricas es independiente del lenguaje, mientras que la aproximación PLN no sólo requiere conocimiento propio de la lengua, sino también herramientas no siempre disponibles para todos los idiomas de la colección.

Aún así, hay idiomas como el alemán, en el que las reglas de composición de palabras hacen prácticamente imposible el acceso a los sintagmas sin una distinción apropiada de las componentes léxicas de los términos.

A pesar de todo esto, y aunque las aproximaciones IR pueden tratar colecciones grandes, no pueden, por el contrario, abordar los problemas de variación terminológica (Jacquemin 2000), lo cual termina por afectar a la calidad de los sintagmas que se le ofrecen al usuario para su exploración. Por ejemplo, Greenstone (Witten 1999a;Paynter 2001a) ofrece al usuario los sintagmas que contienen una determinada palabra. Si el usuario ha introducido "*bosque*" el

sistema no devolverá ningún sintagma que contenga "*bosques*". Esto se debe a que los sintagmas en sí no se indexan, ni tampoco se normalizan sus componentes, sino que únicamente se trabaja con subcadenas alfanuméricas.

Así pues, abordar los problemas de variación morfosintáctica, semántica y translingüe a través de la consideración de sintagmas requiere la normalización e indexación de sus componentes. Es decir, a partir del conjunto de palabras de la consulta, el sistema debe recuperar sintagmas que contengan no sólo dichas palabras, sino también cualquier variación flexiva de las mismas, sinónimos válidos o traducciones.

4.2.2.1 Normalización de las palabras componentes de un sintagma

La normalización de las componentes de un sintagma para su indexación puede evitar el procesamiento lingüístico (etiquetado de categoría gramatical y lematización) mediante algún procedimiento de *stemming*. Sin embargo, la consideración de variaciones semánticas y translingües de la consulta requiere la utilización de recursos léxicos (diccionarios electrónicos fundamentalmente) cuyas entradas no se corresponden con *stems* o raíces, sino con lemas. Es necesario tener esto en cuenta porque si se opta por algún tipo de *stemming* habrá que modificar correspondientemente los recursos léxicos, sin que ello garantice una mejora en la recuperación de información.

El etiquetado (desambiguación) de categoría gramatical es una tarea costosa que debe aplicarse a todos los documentos de la colección. Esto resulta muy costoso y casi inviable para colecciones tamaño TREC. Como dato, los dos primeros prototipos del sistema implementado trabajaban sobre una colección de unos 1.000 documentos cuyo etiquetado morfosintáctico necesitó varios días de procesamiento. Si bien los ordenadores cada vez son más potentes y las herramientas son más eficientes, el salto a centenas de miles de documentos aún resulta demasiado costoso.

Por esta razón, la extracción de sintagmas, así como la normalización de sus componentes, se realizará únicamente sobre la base de un análisis morfológico de las palabras de la colección. De esta forma se van a integrar la normalización con la extracción en sí de los sintagmas: el análisis morfológico proporciona, además de lemas candidatos, etiquetas morfosintácticas que resultan muy útiles para identificar sintagmas nominales.

Al contrario que el *stemming,* el análisis morfológico no proporciona una única forma base para cada palabra, y por ello será necesario estudiar las

particularidades de la tarea de extracción de sintagmas en el ámbito concreto de la Recuperación de Información para obtener una normalización adecuada de los sintagmas.

4.2.2.2 Extracción de sintagmas mediante patrones morfosintácticos

La extracción de sintagmas, por tanto, se realizará sobre la base de patrones morfosintácticos, tomando como sintagmas adecuados aquellas secuencias de palabras cuyas etiquetas se ajusten a alguno de dichos patrones.

En una primera aproximación, la extracción de sintagmas siguió una metodología propia de la Extracción automática de Terminología (TE) (Bourigault 1992;Frantzi 1999), dirigida a identificar y obtener los términos (palabras o sintagmas) que usualmente se utilizan para referirse a los conceptos de un determinado dominio (ver apartado 5.2, Extracción Automática de Terminología). Sin embargo, una vez estudiado el proceso de extracción automática de terminología se estudiaron sus diferencias con el proceso de indexación y de recuperación de información con el fin de adaptar las técnicas desarrolladas.

En la tarea de Extracción Automática de Terminología, el objetivo es decidir qué términos son relevantes dentro de un determinado dominio. Sin embargo, en la tarea de Recuperación de Información es el usuario quien decide, ya sea escribiéndolos en un área de texto o ya sea de forma interactiva, qué términos son relevantes para la información que busca. Por tanto, el sistema debe conservar el mayor número de términos (palabras y sintagmas) contenidos en la colección. Es decir, teniendo en cuenta la tarea de Recuperación de Información resulta menos determinante la precisión en la extracción de los términos de indexación y más importante asegurar un índice de recuperación elevado.

Las consecuencias de esto son que el proceso de extracción de terminología puede y debe relajarse en los siguientes aspectos:

1. No es necesario establecer criterios de precisión para truncar las listas terminológicas.
2. Los patrones de sintagmas terminológicos pueden relajarse para aumentar el *recall* aún a costa de disminuir la precisión del proceso.

De esta forma, los últimos prototipos del sistema implementado *(Website Term Browser)* han utilizado un único patrón en las cinco lenguas (español, inglés, francés, italiano y catalán) para identificar los sintagmas de la colección. Este patrón viene determinado por la siguiente expresión regular:

*[PHR_CONTENT] [PHR_CLOSED | PHR_CONTENT] * [PHR_CONTENT]*

donde

- *PHR_CONTENT* puede ser un nombre, un adjetivo, un numeral, un infinitivo o un participio. Se trata de las palabras del sintagma con contenido léxico-semántico. Los sintagmas serán indexados por los lemas correspondientes a estas palabras.
- *PHR_CLOSED* puede ser un artículo, un determinante[9], una preposición o una conjunción. Son palabras que participan en el sintagma pero que, por no tener contenido léxico-semántico, no se utilizan para indexar el sintagma.

El resto de etiquetas incluyendo verbos, adverbios, determinantes, signos de puntuación, etc., constituyen delimitadores del sintagma (*BOUND*).

De esta forma, según la expresión regular, un sintagma es cualquier secuencia de palabras etiquetadas de la siguiente forma: la primera y la última palabra son *PHR_CONTENT* y las palabras intermedias son *PHR_CONTENT o PHR_CLOSED* indistintamente.

Este patrón se dirige a la obtención del mayor número de sintagmas aunque se extraigan algunas secuencias incorrectas. Como ya se ha mencionado y se profundizará más adelante, el proceso de recuperación ignorará de forma natural la mayoría de los sintagmas incorrectos.

La detección de este patrón en los textos es muy rápida, pero requiere el etiquetado previo de los textos. Esto, como se ha mencionado, resulta inviable para colecciones grandes debido a su coste computacional. Sin embargo, el patrón utilizado en realidad sólo necesita distinguir tres tipos de palabras. Por ejemplo, la distinción entre nombres y adjetivos no es necesaria. De esta manera, también el etiquetado morfosintáctico podrá relajarse para hacerlo computacionalmente viable.

4.2.2.3 Etiquetado morfosintáctico

El etiquetado que debe realizarse incluye la obtención del lema de cada palabra. Esto ofrece la posibilidad de recuperar sintagmas por sus componentes, independientemente de su forma flexiva. Esta cualidad es importante para cumplir el objetivo de acercar el lenguaje de la consulta a la terminología utilizada en la

[9] La distinción entre artículos y determinantes viene dada por el analizador morfológico utilizado.

colección, pero resulta fundamental para trabajar con diccionarios y poder expandir y traducir las palabras de la consulta.

En los primeros prototipos del sistema implementado, el tamaño de las colecciones permitía un etiquetado completo de los textos. Sin embargo, en el tercer prototipo esta aproximación resultó inviable y se optó por asignar a cada palabra el lema y categoría más frecuentes. Esto permitía acelerar el proceso ya que el etiquetado se realizaba para una sólo ocurrencia de cada palabra. Sin embargo, esta aproximación no estaba exenta de errores y en el cuarto prototipo se decidió estudiar la ambigüedad morfosintáctica dentro de los sintagmas atendiendo a las particularidades de la tarea de extracción de sintagmas.

El etiquetado de categorías proporciona, en general, mucha más información de la que es estrictamente necesaria para identificar un sintagma candidato. Aprovechando la información estadística del sintagma en la colección, es posible realizar un etiquetado menos preciso siempre y cuando los errores de etiquetado no provoquen una pérdida en el número de sintagmas extraídos. Es decir, como se dispone de mecanismos adicionales para pesar la corrección de un sintagma, el etiquetado puede ser menos preciso. Estos mecanismos adicionales, como la consideración de la frecuencia de aparición de un sintagma en la colección, permiten desechar sintagmas terminológicos erróneos que se han extraído a causa de un etiquetado incorrecto.

La extracción de sintagmas requiere un etiquetado de acuerdo con las tres categorías que definen el patrón: *PHR_CONTENT, PHR_CLOSED y BOUND*. La simplificación de la tarea de etiquetado debe garantizar:

1. Que no se pierdan sintagmas aunque se reduzca la precisión de la extracción. Esta premisa es importante porque la recuperación de sintagmas se va a basar en la co-ocurrencia de palabras de la consulta en un mismo sintagma. Esta restricción es muy fuerte, como ya se ha discutido, y es necesario garantizar que no se pierde información.

2. Que se asigne una forma base a cada una de las palabras para poder indexar los sintagmas por sus componentes lematizadas. La forma base elegida debe asegurar que se recuperan los sintagmas independientemente de las formas flexivas utilizadas tanto en los sintagmas como en las consultas.

4.2.2.4 Etiquetado del español y catalán

Para estudiar cómo se puede relajar el etiquetado se han extraído todas las palabras diferentes que contiene la colección CLEF 2001 para el español. Esta colección tiene aproximadamente 1 Gb y 377.969 palabras diferentes que se han considerado en el proceso de indexación. A cada una de estas palabras se le han asignado todas las etiquetas posibles mediante el analizador morfológico MACO (Carmona 1998).

Las palabras se han dividido en los siguientes tipos atendiendo a la salida del analizador morfológico:

1. Palabras que reciben varias etiquetas candidatas pero ninguna corresponde a *PHR_CONTENT o PHR_CLOSED*, es decir, la palabra no tiene ninguna oportunidad de pertenecer a un sintagma.
2. Palabras que reciben varias etiquetas candidatas y todas las etiquetas corresponden a *PHR_CONTENT (*nombre, adjetivo, numeral, infinitivo, participio), *verbo o adverbio.*
3. Palabras que reciben un único lema y categoría
4. Palabras que no reciben categoría y lema alguno.
5. Resto de casos.

Para cada uno de estos tipos se ha definido el siguiente tratamiento.

1. Palabras que no participan en los sintagmas

En el primer caso, puesto que se trata de palabras cuya función únicamente es delimitar los sintagmas, cualquier error de etiquetado o asignación de lemas no afecta en absoluto a la extracción de sintagmas. Estas palabras se etiquetan como *BOUND*.

2. Palabras de categorías abiertas

En el segundo caso, a las palabras ambiguas se les asigna la etiqueta *PHR_CONTENT*. Si la asignación es errónea, es decir, en realidad se trata de un verbo o un adverbio, no se perderán sintagmas sino que, en el peor de los casos, se extraerán sintagmas incorrectos. Por ejemplo, del fragmento de texto *"condicionado a que en el país haya convulsiones",* uno de los sintagmas extraídos es *"país haya*", en el que se ha tomado *haya* como nombre de forma incorrecta. Obsérvese que resulta muy difícil que una consulta se asemeje a esta expresión, por lo que es previsible que la detección incorrecta de este

sintagma no vaya a afectar a la recuperación, simplemente porque no se dará el caso de intentar recuperarlo.

El siguiente problema es asignar el lema correcto a este tipo de palabras. Puesto que un sintagma no contiene verbos o adverbios, los lemas como verbo o adverbio se ignoran incluidos los participios e infinitivos. El error de lematización entonces, se produce por confusión entre el resto de categorías de *PHR_CONTENT*: nombres, adjetivos y numerales.

Una confusión en la lematización puede provocar que una consulta no pueda recuperar algunos sintagmas relevantes. Por ejemplo, *cruces* lematizado en el texto como *cruz* no podrá ser recuperado por una consulta que contenga la palabra *cruce*, porque ninguno de los lemas de *cruce* es el lema *cruz*.

En el caso del español, de las 80.988 palabras diferentes en la colección de CLEF que reciben etiquetas por el analizador morfológico sólo hay 2.473 palabras de este tipo que tengan lemas diferentes. De ellos, 2.082 (84%) sólo se diferencian en la consideración del género, que uno de sus lemas termina en *–o* y el otro en *–a.*. Las restantes 391 palabras (que suponen el 0.5%) tienen una casuística variada que no merece la pena caracterizar.

Estos números muestran que los errores de etiquetado no van a ser muy graves de cara a la extracción y normalización de los sintagmas, y que, en todo caso, puede sistematizarse en el proceso de recuperación. De esta forma, es posible evitar el costoso proceso de etiquetado de los textos.

3. Palabras no ambiguas

En el caso de que el analizador morfológico solo proporcione un lema y categoría, obviamente el etiquetado está exento de errores.

4. Palabras sin etiqueta morfosintáctica

De las 377.969 palabras diferentes encontradas en CLEF 2001, el analizador no ha proporcionado categoría ni lema al 66%, lo que supone un índice muy elevado. El 3% (7.000) corresponden a números que se han querido conservar en la indexación. El resto corresponden a nombres propios, cadenas incorrectas o extrañas y palabras de otros idiomas.

A estas palabras sin etiqueta morfosintáctica se les ha asignado la categoría de *PHR_CONTENT* y se les ha asociado la misma palabra como forma canónica. De esta manera, se respetan las dos premisas de que no se pierdan sintagmas y que se les asignen formas base que permitan su indexación y recuperación.

5. Resto de palabras con ambigüedad morfosintáctica

En el tercer caso entran muy pocas palabras y podrían tratarse caso por caso. Sin embargo, para la desambiguación de estas palabras se ha utilizado el siguiente orden de preferencia heurístico de las etiquetas:

1. Adverbio (R)
2. Determinante (D), artículo (T), preposición (S) o conjunción (C).
3. Numeral (M) o adjetivo (A).
4. Pronombre (P).
5. Nombre (N).
6. Verbo (V).

Una misma palabra puede recibir varias etiquetas diferentes dentro de la misma categoría gramatical. Por ejemplo, una forma verbal que coincide en la primera y tercera persona recibe dos etiquetas verbales diferentes. En la *Tabla 4-1* se detallan todos los tipos de ambigüedad encontrados en el caso 5, mostrando únicamente las etiquetas gramaticales sin considerar sus atributos. Para cada caso de ambigüedad se muestran las palabras a las que afectan en la colección española de CLEF 2001, y la clasificación que produce el orden heurístico de etiquetas respecto a las categorías *PHR_CONTENT, PHR_CLOSED, BOUND*.

Casos de ambigüedad	Palabras en el caso	Clasificación
C V	como	PHR_CLOSED
S V V V	desde entre	PHR_CLOSED
C C R	ya	BOUND
C R	luego entonces aún mientras incluso siquiera	BOUND
C V V V	sea	PHR_CLOSED
C V V	ora	PHR_CLOSED
A D P	algunos alguna algunas	PHR_CLOSED
D N P	ese este ambos cuantos	PHR_CLOSED
A P	último últimas últimos última alguno cualquiera	PHR_CONTENT
M P	primeros segundos	PHR_CONTENT
D N	mi tantos don mis	PHR_CLOSED
D N N	tantas tanta	PHR_CLOSED
N P	cuya tercero cuyas cuyo cómo cuyos te dónde míos quintos tuya tuyas	BOUND
M N P	primera segundos tercera primeras terceras terceros	PHR_CONTENT
P T V	unas	PHR_CLOSED
D N P R	todo	BOUND
N P V V	mía	BOUND
M N P V	segundo quinto	PHR_CONTENT
N P V	uno quintas mías	BOUND
M N P V V	segunda quinta	PHR_CONTENT
N P R V V	nada	BOUND
D N R	tanto cuanto	BOUND
N P R	sí	BOUND
A P R	demás	BOUND
N S	de tras ante so	PHR_CLOSED
C N	si sino pero cuando cuándo conque	PHR_CLOSED
N N S	contra	PHR_CLOSED
M T	un	PHR_CLOSED
N P T	la los las lo	PHR_CLOSED
N S V V V V V	para	PHR_CLOSED
N S V V V	sobre	PHR_CLOSED
A N S V	bajo	PHR_CLOSED
N S V V	cabe	PHR_CLOSED
M N P T V V V	una	PHR_CLOSED
C C	porque	PHR_CLOSED
C P	que donde	PHR_CLOSED
P T	unos	PHR_CLOSED

Tabla 4-1. Casos de ambigüedad tipo 5 en el caso del español

4.2.2.5 Adaptación de los recursos léxicos

Las entradas de los diccionarios son lemas en los que se consideran los acentos. Por esta razón, los primeros prototipos exigían que los usuarios utilizaran acentos apropiadamente en sus consultas. Desgraciadamente los usuarios no se preocupan de la acentuación y ha sido necesario eliminarlos. La normalización de los sintagmas de forma consistente con los recursos léxicos ha obligado a la adaptación de todos los diccionarios incluido EuroWordNet con el fin de sustituir los caracteres especiales.

4.2.2.6 Etiquetado de inglés, francés e italiano

Al igual que el analizador morfológico, las heurísticas que pueden mejorar la eficiencia del etiquetado morfosintáctico con el fin de extraer sintagmas son dependientes del idioma. El inglés es un idioma mucho menos flexivo que el español y la obtención de formas base es sencilla. De hecho, los algoritmos de *stemming* para el inglés proporcionan muy buenos resultados y, en general, se acepta que esta técnica permite mejorar la recuperación.

Puesto que no disponemos de un analizador morfológico para el inglés, francés e italiano, la forma de asignar el lema y la categoría gramatical a cada palabra se basa en la herramienta de etiquetado utilizada (TreeTagger). Se trata de un etiquetador supervisado con la opción de asignar a las palabras el lema y categoría más frecuentes en el corpus de entrenamiento. Las palabras en inglés, francés e italiano se han etiquetado utilizando esta opción.

En el caso del inglés, los errores de lematización no son demasiado graves porque se trata de una lengua poco flexiva. Sin embargo, en el caso del francés y el italiano los errores de etiquetado se reflejan en una pérdida de *recall* respecto al español, el inglés y el catalán. Esto evidencia la fuerte dependencia de la lengua que tiene el empleo de conocimiento, técnicas y recursos lingüísticos en el acceso a la información.

4.2.2.7 Aplicación de los patrones morfosintácticos

El resultado de aplicar los patrones morfosintácticos es la extracción de los sintagmas cuyas componentes se ajustan al patrón. Pero, además, a cada sintagma se le asocian los lemas de sus componentes con contenido semántico (*PHR_CONTENT* –adjetivos y nombres principalmente). Estos lemas permitirán indexar y recuperar los sintagmas independientemente de la forma flexiva en que se utilicen tanto en la consulta como en los documentos.

Estos patrones morfosintácticos se aplican exhaustivamente sobre los textos permitiendo la identificación de sintagmas que se solapan o contienen entre sí. Las ventajas de esta aproximación son que no se pierden posibles candidatos y que su procesamiento es muy rápido debido a su sencillez. La contrapartida es el volumen de sintagmas que llega a extraerse y que habría que considerar en la recuperación de los términos. Sin embargo, es posible descartar un porcentaje muy elevado de sintagmas si se consideran relaciones de subsunción entre sintagmas así como la información estadística asociada a cada sintagma tras el proceso de indexación de los documentos.

4.2.3 Indexación de los documentos

Se realizan dos indexaciones de los documentos. Una en la que los documentos están indexados por sus lemas, y otra en la que están indexados por sus sintagmas. La primera indexación es la usual en Recuperación de Información utilizando los lemas en lugar de las palabras originales. La segunda es análoga sólo que se utilizan los sintagmas como términos de indexación de los documentos.

La única particularidad de la indexación mediante sintagmas es que debe considerar un número de términos mayor en varios órdenes de magnitud. Esto, en colecciones grandes, hace necesaria la aplicación de algoritmos de construcción de ficheros invertidos atendiendo a restricciones de memoria. Además, hace necesario algún tipo de selección de sintagmas.

4.2.4 Selección de sintagmas

Todo el proceso descrito hasta el momento se ha realizado bajo la premisa de no perder sintagmas que puedan ser relevantes. Esto, dependiendo del tamaño de la colección, hace que el número de sintagmas extraídos pueda ser inviable desde un punto de vista computacional.

Sin embargo, la consideración de sintagmas en la colección proporciona algunos criterios que permiten reducir el número de sintagmas candidatos en más de un 75%. Esta reducción es porcentualmente mayor cuanto mayores son las colecciones. Por ejemplo, en la colección española de CLEF 2001 de casi 1 Gb, la reducción del número de sintagmas candidatos fue del 85%.

A continuación se describen los criterios utilizados para la selección de sintagmas.

4.2.4.1 Subsunción de sintagmas

Como el ajuste de patrones se realiza de forma exhaustiva permitiendo el solapamiento de sintagmas extraídos de un mismo fragmento, muchos de los sintagmas candidatos son subsintagmas de otros más largos. Debido a que la indexación de los sintagmas se realiza por los lemas de sus componentes abiertas (nombres y adjetivos), consideramos que el sintagma i contiene al sintagma j si el conjunto de las componentes del sintagma i incluye al conjunto de las componentes del sintagma j.

La decisión de si los sub-sintagmas deben conservarse o no, debe hacerse depender de la colección. Para ello definiremos una relación de subsunción de la siguiente manera:

un sintagma i subsume a un sintagma j si j es una subcadena de i contenida en los mismos documentos que j

Los sintagmas subsumidos en otro sintagma podrán descartarse sin pérdida de información. Por ejemplo, considérense los siguientes sintagmas:

	cod_sintagma	sintagma	df.	frec.
1	299657	alumnos de la facultad	64	122
2	320195	alumnos de la facultad de derecho	2	2
3	320196	alumnos de la facultad de derecho de la uned	2	2
4	1216399	facultad de derecho	714	883
5	1217129	facultad de derecho de la uned	19	25
6	666584	derecho de la uned	22	29

El sintagma más largo que contiene a los demás es el sintagma (3). Si se conserva este sintagma descartando el resto, entonces se perdería el acceso a los documentos indexados por el resto de sintagmas. El caso extremo es el sintagma (4) que indexa 714 documentos diferentes mientras que el sintagma (3) sólo permite acceder a 2 de ellos.

Sin embargo, el sintagma (3) también contiene al sintagma (2) y, en este caso, ambos referencian a los mismos documentos (aunque en la tabla sólo se muestra el número de documentos). Como el sintagma (3) se indexa por los mismos lemas que (2) (*alumno, facultad, derecho*) y por un lema más (*uned*), todas las consultas que recuperen el sintagma (2) también recuperarán el sintagma (3), mientras que al revés no tiene por que ocurrir (v.g. aparezca *uned* en la consulta). Así pues, el sintagma (2) está subsumido por el sintagma (3) y se puede eliminar sin pérdida alguna de información. En el ejemplo, no se podría descartar ningún otro sintagma.

Debe observarse que, de esta manera, la decisión de descartar un sintagma requiere que se haya realizado previamente la indexación de la colección por sintagmas.

La detección de todos los posibles casos de subsunción entre los sintagmas extraídos de una colección puede hacerse computacionalmente muy costosa. La forma de abordar el problema ha sido la siguiente. Para que haya subsunción entre sintagmas, al menos debe producirse en un documento. Así pues, si en cada documento se extraen de forma exhaustiva todos los sintagmas, incluidos los casos de subsunción, para detectar la subsunción entre sintagmas basta con comparar los sintagmas extraídos de un mismo documento. Esto reduce considerablemente el problema, pero aún puede simplificarse más si la extracción de sintagmas se hace de forma ordenada considerando la relación de subsunción de forma implícita. Considérese el siguiente fragmento:

pedagogía social y sociología de la educación

Todas los subsintagmas posibles de este fragmento son:

(1) *pedagogía social*
(2) *pedagogía social y sociología*
(3) *pedagogía social y sociología de la educación*
(4) *social y sociología*
(5) *social y sociología de la educación*
(6) *sociología de la educación*

En primer lugar, no es posible que exista otro documento que contenga sub-sintagmas de este fragmento que no se hayan contemplado ya. Sí puede ocurrir que exista otro documento con un sintagma más largo que contenga este fragmento completo. En ese caso, el razonamiento se aplicaría de igual modo para aquel documento, ya que contendría todos los casos posibles de subsunción incluidos los de nuestro ejemplo.

En segundo lugar, el orden de extracción de los sintagmas de nuestro ejemplo permite seguir un esquema muy sencillo por comparación de un subsintagma con el siguiente en la lista:

1. (1) está contenido en (2)
2. (2) está contenido en (3)
3. (3) contiene a (4) (5) y (6)
4. (4) está contenido en (5)
5. (5) contiene a (6)
6. (6) no participa en más relaciones

Apoyándose en este esquema, el algoritmo de selección de sintagmas basado en subsunción es el siguiente:

```
Bucle
   Si SINTAGMA_ACTUAL está_contenido_en SINTAGMA_SIGUIENTE
   Entonces
        Si LISTA_DOCS_SINTAGMA_ACTUAL está_contenida_en LISTA_DOCS_SINTAGMA_SIGUIENTE
        Entonces
                Descartar SINTAGMA_ACTUAL
        Fin Si
        SINTAGMA_ACTUAL ← SINTAGMA_SIGUIENTE
   Si no
        Si SINTAGMA_ACTUAL contiene_a SINTAGMA_SIGUIENTE
        Entonces
                Si LISTA_DOCS_SINTAGMA_ACTUAL contiene_a LISTA_DOCS_SINTAGMA_SIGUIENTE
                Entonces
                        Descartar SINTAGMA_SIGUIENTE
                Fin Si
        Si no
                SINTAGMA_ACTUAL ← SINTAGMA_SIGUIENTE
        Fin Si
   Fin Si
Fin Bucle
```

La lista de documentos asociada a un sintagma es, por construcción, una lista ordenada de números, por lo que la decisión de si un sintagma se descarta o no es de coste lineal con respecto al número de documentos que contienen al sintagma, que puede llegar a ser muy elevado. Sin embargo, el coste se puede hacer constante si redefinimos la relación de subsunción como:

el sintagma i subsume al sintagma j si el sintagma j es una subcadena del sintagma i que aparece en el mismo número de documentos que j

Si un sintagma i es un super-sintagma de un sintagma j (j es una subcadena de i), entonces el sintagma i aparece al menos en todos los documentos en los que aparece el sintagma j, lo que implica que el número de documentos en los que aparece i es menor o igual que el número de documentos en los que aparece j, y que si es igual, ambos aparecen en los mismos documentos.

En conclusión, si el sintagma i contiene al sintagma j, y ambos aparecen en el mismo número de documentos, entonces el sintagma j se puede descartar pues contiene menor número de lemas y sin embargo está referenciando los mismos documentos. El algoritmo queda entonces como:

```
Bucle
   Si SINTAGMA_ACTUAL está_contenido_en SINTAGMA_SIGUIENTE
   Entonces
        Si NUM_DOCS_SINTAGMA_ACTUAL = NUM_DOCS_SINTAGMA_SIGUIENTE
        Entonces
                Descartar SINTAGMA_ACTUAL
        Fin Si
        SINTAGMA_ACTUAL ← SINTAGMA_SIGUIENTE
   Si no
        Si SINTAGMA_ACTUAL contiene_a SINTAGMA_SIGUIENTE
        Entonces
                Si NUM_DOCS_SINTAGMA_ACTUAL = NUM_DOCS_SINTAGMA_SIGUIENTE
                Entonces
                        Descartar SINTAGMA_SIGUIENTE
                Fin Si
        Si no
                SINTAGMA_ACTUAL ← SINTAGMA_SIGUIENTE
        Fin Si
   Fin Si
Fin Bucle
```

De esta forma, el algoritmo resultante descarta sintagmas bajo el criterio de subsunción con un coste lineal sobre el número de sintagmas extraído de la colección. Esto es así porque únicamente realiza una evaluación por sintagma para decidir si se descarta o no, sólo se evalúa el sintagma con respecto al siguiente sintagma extraído del documento y el coste de la comparación es constante. El algoritmo, aplicado al ejemplo:

	cod_sintagma	sintagma	df	frec.
1	1855070	pedagogía social	97	135
2	1875655	pedagogía social y sociología	23	25
3	1875656	pedagogía social y sociología de la educación	23	25
4	2236613	social y sociología	23	25
5	2236614	social y sociología de la educación	23	25
6	2218148	sociología de la educación	56	84

permite descartar (2), (4) y (5) dejando como sintagmas relevantes para la recuperación (1), (3) y (6).

4.2.4.2 Grado de lexicalización

El grado de lexicalización de un sintagma es un criterio menos preciso que el criterio de subsunción, pero que permite descartar también gran número de sintagmas. Una expresión lexicalizada debe aparecer como tal con una determinada frecuencia. Si la colección es suficientemente grande, una expresión que únicamente aparece una vez tiene muchas posibilidades de no ser una expresión

lexicalizada. Como sintagma terminológico tiene escaso valor ya que no representa un concepto por si misma. Así pues, de acuerdo con este criterio, se pueden seleccionar todos los sintagmas que aparezcan más de una vez en la colección.

Puede ocurrir que un documento utilice varias veces una expresión sin que dicha expresión suponga la lexicalización de un concepto. La probabilidad de que esto ocurra considerando documentos diferentes disminuye notablemente. Así pues, se puede fortalecer la estimación del grado de lexicalización no ya mediante la frecuencia del sintagma en la colección, sino por el número de documentos diferentes que lo contienen. De esta forma, se seleccionan todos los sintagmas que aparezcan en más de un documento de la colección.

La aproximación del grado de lexicalización por el número de documentos en los que aparece se comporta mejor cuanto mayor sea la colección. Sin embargo, en general una colección nunca será suficientemente grande como para que todas las expresiones lexicalizadas aparezcan más de una vez por lo que algunas de ellas se descartarán erróneamente. Al contrario que en el criterio de subsunción, con el criterio de conceptualización sí se pueden perder expresiones relevantes para la búsqueda. No sólo expresiones que sí deberían considerarse como lexicalizadas, sino también sintagmas que, aún no siendo lexicalizaciones, relacionaban términos de búsqueda de forma útil.

A pesar de ello, la heurística tiene un coste computacional muy bajo y la mayoría de los sintagmas descartados resultan ser expresiones erróneas y expresiones no lexicalizadas, por lo que, en general, es una heurística que mejora la calidad de los sintagmas que se ofrecen al usuario.

En la colección española de CLEF 2001 se ha aplicado este criterio de selección de sintagmas. El proceso de extracción ha proporcionado aproximadamente 26,7 millones de sintagmas candidatos. De ellos sólo 3,6 millones aparecen en más de un documento de la colección (aproximadamente el 15%). La reducción pues, es muy significativa y puede ser determinante para que la indexación de sintagmas sea viable en colecciones tan grandes.

4.2.5 Proceso de indexación

En resumen, el proceso de extracción de sintagmas e indexación resultante es el siguiente:

1. Preprocesamiento del texto: eliminación de etiquetas de marcado, tokenización y detección del idioma.
2. Extracción de las palabras de la colección (sólo una ocurrencia).
3. Etiquetado de categoría gramatical y asignación de forma base.
4. Detección de sintagmas candidatos mediante la aplicación de patrones morfosintácticos sobre las etiquetas de categoría. Asignación al sintagma de los lemas que contiene.
5. Indexación de documentos por los términos (lemas y sintagmas) que contiene. Obtención de las frecuencias de aparición de los términos en la colección.
6. Selección de sintagmas sobre la base de las estadísticas obtenidas.
7. Indexación de los sintagmas seleccionados por sus lemas componentes.

Sólo el paso 3 de etiquetado resulta dependiente de la lengua, si bien la consideración de nuevas lenguas puede llevar a la consideración de patrones específicos en el paso 4.

La *Figura 4-1* muestra el esquema de indexación en el que los documentos se indexan por sintagmas y lemas, y a su vez los sintagmas se indexan por sus lemas.

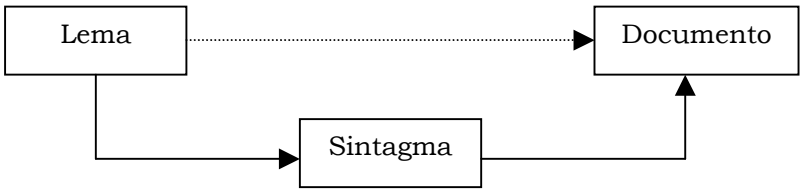


Figura 4-1. Esquema de indexación

4.3 Modelo propuesto de recuperación

Como ya se ha mencionado, el coste de considerar sintagmas es una disminución en el índice de recuperación (*recall*). El índice de recuperación puede elevarse mediante la expansión de la consulta, es decir, añadiendo términos a la consulta de

forma automática. En este trabajo, el caso de traducción puede contemplarse como un caso particular de expansión de la consulta en el que la única diferencia es que los términos que se añaden a la consulta son de otro idioma.

Cuando se traduce o se expande una consulta palabra por palabra, la ambigüedad de las palabras aisladas hace que la expansión añada ruido a la consulta disminuyendo la precisión de la recuperación. Para resolver este problema existen varias estrategias ya sea mediante análisis local o análisis global. La estrategia más común es tratar de seleccionar los términos de expansión que sean similares no ya a las palabras aisladas de la consulta sino a varias o a todas ellas en conjunto (Mandala 1999).

Para establecer una medida de similitud entre los términos de traducción o expansión candidatos y las palabras de la consulta, en la literatura se consideran tanto estadísticas de co-ocurrencia diversas sobre sintagmas y colocaciones, como correlación de patrones de co-ocurrencia, características contextuales, etc.

Estas estrategias se combinan con el uso de tesauros ya sean manuales, automáticos (matrices de co-ocurrencia) o combinaciones de ambos tipos.

En nuestro caso, la utilización de sintagmas para reducir la ambigüedad se realiza de forma similar a (Ballesteros 1998). Estos autores traducen palabra a palabra sintagmas no contemplados en el diccionario. Su hipótesis es que las traducciones correctas de un sintagma co-ocurren con frecuencia en la colección mientras que las incorrectas no. De esta manera puede obtenerse una medida de posibilidad para la traducción de colocaciones y sintagmas. Cada una de las palabras del sintagma tiene varias traducciones, dando varias combinaciones posibles como traducción del sintagma. Se selecciona aquella combinación que aparece más veces en la colección.

La *Figura 4-2* muestra un ejemplo de expansión por sinónimos así como de traducción mediante EuroWordNet. En el ejemplo puede observarse el problema que origina la ambigüedad de las palabras en la traducción. ¿Cuál es la mejor traducción para la consulta *"Tratados de prohibición de pruebas nucleares"*? En este caso, *"Nuclear test ban treaty"* tiene mayor frecuencia de aparición en la colección de documentos ingleses que el resto de combinaciones de traducción posibles. Este criterio permite, en esta ocasión, seleccionar la traducción más adecuada.

En nuestro trabajo se utiliza una aproximación similar para reducir la ambigüedad no sólo de la traducción, sino también de la expansión de la consulta. La diferencia fundamental es que (Ballesteros 1998) utiliza este método para obtener una única traducción de un sintagma previamente identificado. En nuestro caso, el objetivo es

diferente, basta con identificar alguna combinación de palabras de expansión y de traducción que estén presente en la colección. Es decir, se van a recuperar todos los sintagmas que contengan alguna combinación de las palabras expandidas y traducidas de la consulta.

Consulta	Tratados de	Prohibición de	Pruebas	Nucleares
Expansión	acuerdo capitulación concertación convenio cuidar, pacto manejar procesar	embargo entredicho interdicción interdicto proscripción	cata, catadura degustación ensayo escandallo experimento gustación muestreo, tanteo	nuclear
Traducción	accord discourse handle manage pact process treat treatise treaty	ban interdiction prohibition proscription	demonstrate establish, exhibit experiment experimentation fall, fitting indicate, point present, proof prove, run sample, sampling shew, show, taste test, trial, try	nuclear
Consulta Traducida	Nuclear fitting interdiction manage? Nuclear taste proscription process? Nuclear test ban treaty?			

Figura 4-2. Ambigüedad en la expansión y traducción de la consulta.

El hecho de que algunas de las traducciones y expansiones de las palabras de la consulta co-ocurran en un sintagma indica una relación de interés para el usuario. Además estos sintagmas están directamente relacionados con documentos de la colección puesto que son sintagmas extraídos previamente de la misma durante el proceso de indexación, y van a permitir acceder de forma directa a documentos relevantes para el usuario.

También resulta interesante identificar las palabras aislados que participan en alguna combinación así como aquellas que no participan en ninguna. Atendiendo al criterio de *dispersión léxica* (Anick 1999), aquellas palabras aisladas que aparezcan en un mayor número de sintagmas relacionados con la consulta tienen mayor probabilidad de ser la traducción adecuada de la palabra original.

De esta forma, se recuperan todos los términos (palabras y sintagmas) relacionados con la consulta. Una vez recuperados los términos relacionados con la consulta deben organizarse para que resulten de utilidad para el usuario.

El proceso de recuperación atendiendo al procesamiento de la consulta descrito, viene descrito por la *Figura 4-3.*

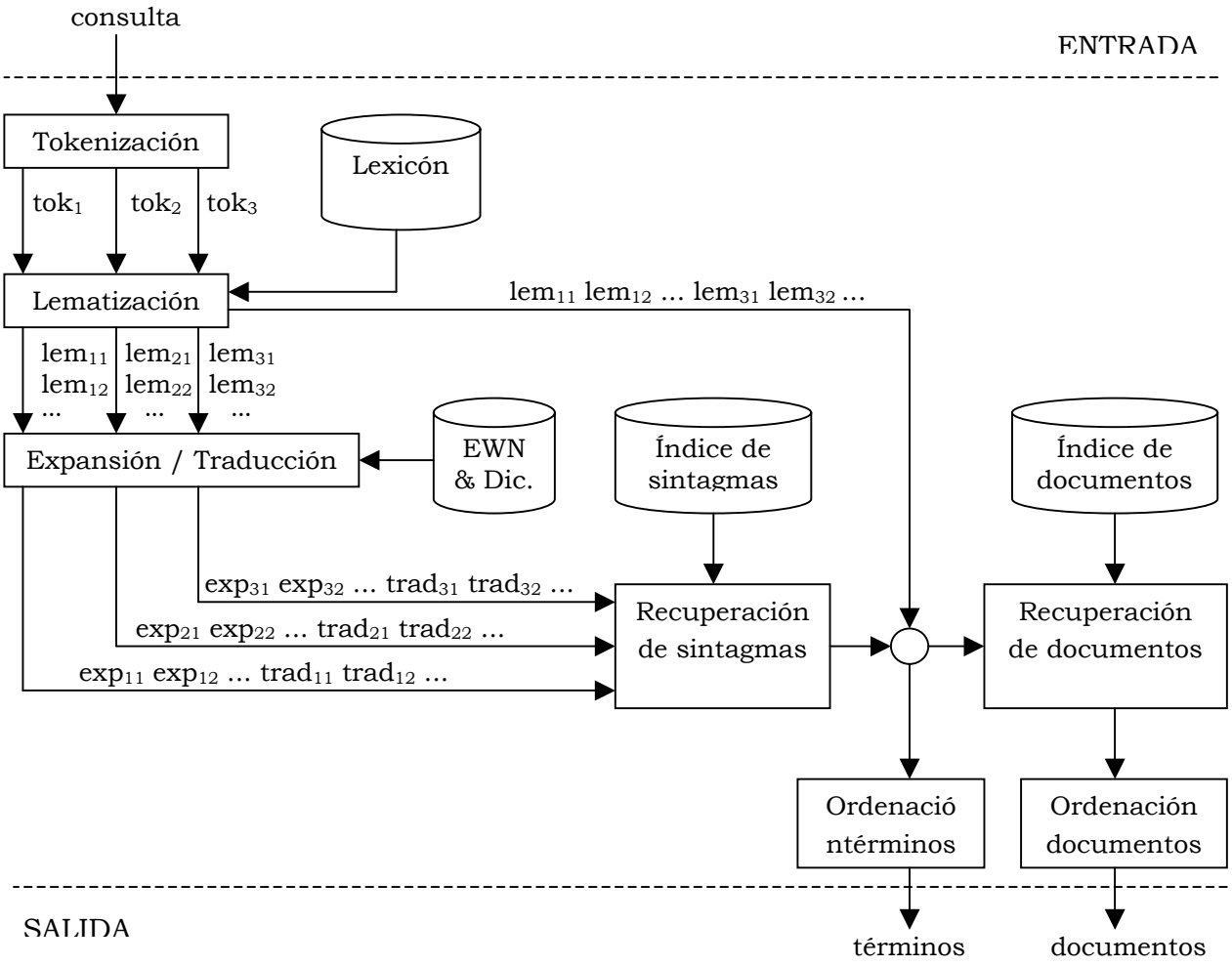


Figura 4-3. Proceso de recuperación.

En primer lugar, se identifican las palabras de la consulta (tokenización), y se hallan todos sus lemas posibles. Cada uno de estos lemas tiene asociados una lista de sinónimos y otra lista de traducciones posibles según los diccionarios y EuroWordNet. Los sinónimos se combinan entre sí para recuperar los sintagmas que los contienen. Análogamente, se combinan las palabras de traducción para recuperar los sintagmas en cada idioma destino que las contienen. Este proceso de recuperación de sintagmas da como resultado una larga lista que hay que ordenar y

organizar antes de ofrecérsela al usuario en el proceso de interacción. Por último, todos los términos (palabras y sintagmas en el mismo y otros idiomas) son considerados en la recuperación de documentos para ofrecer al usuario un ranking inicial de documentos.

A continuación se detalla cada una de las fases del proceso.

4.3.1 Consulta

La consulta al sistema se realiza mediante una serie de palabras que introduce el usuario en un área de texto para acotar el objeto de su búsqueda. No es necesario que la consulta tenga estructura sintáctica alguna, aunque el sistema presupone que el usuario ha introducido las palabras clave de un sintagma que describe el objeto de su búsqueda.

El usuario no debe preocuparse de utilizar sinónimos o de ajustar su consulta a la terminología de la colección. Tampoco debe indicar al sistema el idioma de su consulta. El sistema realizará las inferencias que permiten aproximar la consulta a los términos de la colección.

4.3.2 Preprocesamiento y lematización

En primer lugar, se *tokeniza* la consulta eliminando las palabras de categorías cerradas que no indexan ni sintagmas ni documentos. A continuación se obtienen todos los posibles lemas de los tokens aislados. A diferencia de la tarea de extracción de sintagmas, en la consulta sí es necesario considerar los lemas correspondientes a categorías como verbos y adverbios que, si bien no intervienen en los sintagmas, sí son elementos de indexación de documentos.

Por esta razón, en general habrá varios lemas asociados a cada token. En esta ocasión, de nuevo, se sigue la filosofía general de la propuesta, no desambiguar a priori cuando no hay información suficiente, sino dejar que durante la recuperación los lemas apropiados de cada token co-ocurran en los sintagmas de la colección y emerjan de forma natural. El resultado del preprocesamiento es, en general, una lista de lemas para cada uno de los tokens de la consulta. De esta forma, además, cada uno de los lemas diferentes permitirá acceder a distintos sinónimos y traducciones.

Como en tiempo de consulta el uso de un analizador morfológico puede suponer una carga computacional elevada, se ha optado por utilizar un lexicón en forma de base de datos, que proporciona, para cada palabra, todos sus lemas posibles. Esta opción sólo ha sido posible en el caso del español por no disponer de los recursos léxicos apropiados para el resto de lenguas. El lexicón tiene 839.796 entradas por lo que cubre bastante bien el español.

4.3.3 Expansión y traducción de la consulta

El resultado del *preprocesamiento* es una lista de lemas por cada token original. Esta filosofía se conserva en todo el procesamiento de la consulta previo a la recuperación de sintagmas y documentos. En esta ocasión, cada lema se traduce y expande añadiendo nuevas listas de lemas a cada token. Así, además de la lista de lemas originales, a cada token se le asocian las listas de lemas correspondientes a sinónimos y traducciones en español, inglés, catalán, francés e italiano.

Estos lemas de diferentes idiomas se obtienen mediante la base de datos léxica EuroWordNet (Vossen 1998) y mediante diccionarios bilingües. Ha sido necesario adecuar EWN a las particularidades del sistema, tanto en su conversión al sistema gestor de bases de datos que se utiliza como respecto a la eliminación de acentos.

4.3.4 Recuperación y ordenación de sintagmas

La recuperación de sintagmas se realiza por sus términos de indexación, es decir, por los lemas que contienen. De esta forma, se obtienen los sintagmas que contienen alguno de los lemas resultantes de la expansión y traducción. El número de lemas de expansión suele ser bastante elevado, y la utilización de términos relacionados semánticamente (como sinónimos o merónimos) produce mucho ruido en la expansión. Sin embargo, el ranking de sintagmas permitirá descartar la mayoría de los sintagmas irrelevantes fruto de combinaciones inapropiadas de lemas expandidos.

A diferencia de la recuperación translingüe en *batch,* donde la información sintagmática se utiliza únicamente para seleccionar la mejor traducción de acuerdo con su contexto, en este proceso todos los sintagmas obtenidos se conservan para el proceso de selección interactivo. Sin embargo, la gran cantidad de sintagmas obliga a su ordenación de acuerdo con criterios de relevancia como son:

- Proximidad a la consulta: el número de palabras originales, expandidas o traducidas que contiene el sintagma.
- Grado de lexicalización del sintagma: estimado como el número de documentos de la colección en los que aparece el sintagma (*document frequency*).

La forma en que se van a mostrar los sintagmas al usuario, así como los criterios de relevancia de un sintagma, no sólo determinan el ranking, sino también la forma en la que se hace la recuperación de los términos. En WTB se han explorado dos aproximaciones bastante diferentes.

4.3.4.1 Recuperación y ordenación basada en el número de co-ocurrencias y grado de lexicalización

Los sintagmas que contienen mayor número de palabras de la consulta señalan con mayor probabilidad documentos relevantes para el usuario. Como es natural, hay pocos sintagmas que contienen más de dos palabras de la consulta aun considerando los lemas de expansión y traducción. Sin embargo, si existe alguno, éste conducirá al usuario a documentos relevantes casi con total seguridad. El primer criterio de ordenación, entonces es el número de lemas relacionados con la consulta presentes entre las componentes del sintagma.

La mayoría de los sintagmas relevantes tienen una o dos componentes relacionadas con la consulta. En el caso de que haya sintagmas con dos componentes relevantes, los sintagmas se agrupan por dichas componentes formando grupos de sintagmas relacionados entre sí en un segundo nivel de la jerarquía. El representante de un grupo es el sintagma con mayor grado de lexicalización. Tanto el orden de los grupos como de los sintagmas dentro de cada grupo se establece de acuerdo con este criterio.

4.3.4.2 Recuperación y ordenación basada en proximidad a la consulta

Si tenemos en cuenta la presunción de que el usuario ha utilizado la expresión óptima para encontrar la información que busca, el primer intento de recuperación debe utilizar esa expresión sin realizar ningún procesamiento adicional como traducción o expansión de alguna de sus palabras. Es necesario justificar cualquier procesamiento adicional que añada coste a la recuperación. La justificación es que la expresión original no haya recuperado términos relevantes en la colección. En ese caso se requiere procesar la consulta y realizar las inferencias necesarias para poder contextualizarla.

Bajo el supuesto de que la expresión del usuario es óptima, las transformaciones deben realizarse de forma que primero se obtengan las expresiones más cercanas a la forma elegida por el usuario. Como los sintagmas se indexan por los lemas que contienen y estos lemas vienen determinados por la consulta, la forma de alterar la recuperación de sintagmas es modificar alguno de los lemas de la consulta, intercambiándolo con algún sinónimo o alguna traducción. Estas transformaciones se van a agrupar por niveles de acuerdo a la siguiente heurística:

> *Se realizan primero las transformaciones que llevan a recuperan primero los sintagmas con más peso según la fórmula:*
>
> $$2 \cdot lemas\_originales + lemas\_transformados$$

Así, los sintagmas que contienen entre sus componentes todos los lemas originales de la consulta, forman el primer nivel. El segundo nivel lo forman todos los sintagmas que contienen todos los lemas originales de la consulta salvo un sinónimo, etc. La heurística lleva, por ejemplo, a que se intenten recuperar antes los sintagmas con 2 lemas originales que los sintagmas que contienen 3 sinónimos o traducciones de lemas originales.

Dentro de cada nivel, los sintagmas se ordenan por su estimación de grado lexicalización sobre la base del número de documentos en los que aparece.

Este ranking de términos prima la recuperación monolingüe puesto que los sintagmas en otros idiomas son los más alejados de la consulta. Esto, en general, resulta conveniente para los usuarios pero, sin embargo, premiar el idioma sobre la calidad de la terminología recuperada puede llevar a que información más relevante no sea estimada apropiadamente por el usuario.

4.3.4.3 Comparación entre las dos aproximaciones de recuperación y ordenación

La ventaja del ranking basado en proximidad a la consulta respecto al ranking basado en co-ocurrencias y grado de lexicalización es que se establece un criterio de ordenación para todos los sintagmas y que los sintagmas más cercanos a la consulta se van a recuperar con mayor rapidez.

La desventaja, sin embargo, es que los términos no se agrupan por semejanza y, por tanto, la atención del usuario no se puede centrar en grupos de terminología, sino que debe recorrer secuencialmente todos los sintagmas que le ofrece el sistema.

Otra diferencia es que la proximidad a la consulta no garantiza la recuperación en primer lugar de términos con mayor grado de lexicalización. De hecho, es frecuente recuperar en los primeros lugares sintagmas muy largos que acotan de forma muy concreta la información y que únicamente conducen a uno o dos documentos. Esto resulta útil en algunos casos pero, en otros, puede llevar a la recuperación de sintagmas incorrectos en los primeros lugares del ranking, suceso que se evita cuando se garantiza un determinado grado de lexicalización.

4.3.4.4 Desambiguación léxica

La recuperación de sintagmas por sus lemas componentes así como su ordenación atendiendo al número de componentes producen implícitamente una desambiguación léxica. De todas las posibles combinaciones entre los lemas de expansión y traducción, sólo afloran las que, por aparecer en las colecciones, pueden considerarse adecuadas.

Esta desambiguación se realiza tanto en el nivel morfosintáctico como semántico. En el nivel morfosintáctico hay que recordar que a cada palabra original de la consulta se le ha asociado una lista de posibles lemas sin descartar ninguno. La co-ocurrencia de lemas de indexación en un mismo sintagma permite considerar las formas adecuadas y descartar de forma implícita las que no lo son. Por ejemplo, los siguientes casos tienen la misma indexación:

base de datos bibliográfica ⇔ bases de datos bibliográficas
recuperación de información multilingüe ⇔ recuperación multilingüe de la información

En el nivel semántico ocurre lo mismo. A cada lema original de la consulta se le ha asociado una lista de sinónimos. La co-ocurrencia de lemas en un sintagma permite aceptar las combinaciones de sinónimos adecuados, descartando aquellos que quedan fuera de lugar en el contexto. Ocurre lo mismo en el caso translingüe con las traducciones asociadas a cada lema original de la consulta.

4.3.5 Recuperación y ranking de documentos

El sistema indexa los documentos tanto mediante lemas como sintagmas. Por tanto, como es natural, la consulta también recupera documentos que deben ser ordenados de acuerdo con algún criterio de relevancia. Al margen de los rankings tradicionales basados en la frecuencia de las palabras en los documentos y en la colección, la recuperación de terminología asociada a la consulta permite la exploración de nuevos criterios de ordenación de documentos.

Por ejemplo, si consideramos la consulta como la expresión de un concepto central de un dominio, entonces el ranking de documentos se puede realizar atendiendo al criterio de que los documentos relevantes son aquellos que constituyen un contexto en el que la consulta es el concepto central. Este criterio se puede aproximar con el número de términos diferentes contenidos en cada documento y recuperados en el proceso de recuperación de terminología. Es decir, se presentan en primer lugar los documentos que contienen más términos diferentes relacionados con la consulta. Evidentemente muchos de estos términos no contendrán los lemas originales de la consulta, sino que habrán sido resultado de expansión, traducción y combinación de los mismos.

La ventaja de este ranking es que se puede describir un documento en el ranking mediante el conjunto de términos que contiene y que se relacionan con la consulta. Se trata de una descripción que resume todo el contenido relevante del documento aunque de una forma muy escueta. Por ejemplo, en el tercer prototipo, las tres primeras entradas en el ranking de documentos recuperados en la colección de *noticias internacionales* tras la consulta *"tratados de prohibición de pruebas nucleares"* están recogidas en la *Figura 4-4.*

10808 cols/washington/000465.htm
Terms: accord / pact / treaty / ban / experiment / indicate / point / prove / show / test / try / nuclear / ban pact / test ban pact / test ban pact / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / comprehensive test ban / test ban treaty / test ban pact / nuclear test / nuclear tests / nuclear test explosions /

10715 cols/washington/000372.htm
Terms: pact / treaty / ban / point / present / prove / show / test / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / test ban treaty / nuclear test ban / nuclear test ban / nuclear test / nuclear tests / comprehensive nuclear test / first nuclear test / simulated nuclear tests / nuclear test ban /

2577 cols/pais/002581.htm
Terms: acuerdo / tratado / tratar / prohibición / ensayo / mostrar / prueba / nuclear / tratado de prohibición / prohibición de las pruebas / prohibición de las pruebas nucleares / prohibición de pruebas / prohibición de pruebas nucleares / prohibición de las pruebas nucleares / prohibición de pruebas nucleares / ensayos nucleares / realización de los ensayos nucleares / pruebas nucleares / prohibición de las pruebas

Figura 4-4. Ranking y descripción de documentos por sus términos (palabras y sintagmas) relacionados con la consulta.

Puede observarse que el ranking se compone de documentos en cualquiera de los idiomas de la colección (e.g. los dos primeros son en inglés y el tercero en español), lo que supone una aproximación posible al problema no resuelto de fundir rankings en sistemas de recuperación multilingüe. Además, la lista de términos, y en especial los sintagmas, contenidos en el documento y relacionados con la consulta se convierten en buenos elementos de descripción del documento y permiten al usuario determinar la relevancia del mismo.

4.4 Modelo propuesto de interacción

El desarrollo de modelos avanzados de acceso a la información requiere la consideración de los principios que rigen la comunicación ya que los usuarios podrán comunicar al sistema sus objetivos de búsqueda en la medida que las interfaces así se lo permitan. En la mayoría de los buscadores, las interfaces no han evolucionado más allá de un área en la que el usuario enumera una lista de palabras clave. Sin embargo, parece difícil superar este modelo tan sencillo. La presencia de las palabras de la consulta en un mismo contexto (un documento, generalmente) acota de forma implícita el significado de la consulta y, por ende, cada una de sus palabras. Cualquier intento de explicitar este significado de forma previa a la recuperación de los documentos supone, por lo general, una pérdida de información o la introducción de ruido en la búsqueda. La mejor interpretación de una consulta es, en realidad, un documento relevante.

El usuario construye su consulta optimizando el coste de su producción y, por ello, una consulta, por muy rico que sea el lenguaje de la consulta, no será, por lo general, completamente explícita, sino que utilizará un conjunto de supuestos implícitos relativos al conocimiento que el usuario tiene sobre el sistema, sobre su funcionamiento, sobre el contenido de las colecciones, y también a su conocimiento del mundo.

Aunque a veces de forma inconsciente, se maneja el supuesto de que los usuarios conocen el funcionamiento de los sistemas de búsqueda. Esto en general no es cierto y es una de las causas de insatisfacción de los usuarios. Piénsese en lo poco intuitivo que puede ser un modelo sencillo como el modelo de espacio vectorial a la hora de explicar por qué el sistema devuelve un determinado ranking de documentos.

Pero también se asume que los usuarios conocen el dominio de búsqueda y, esto, aunque pueda ser cierto en términos generales, no lo es siempre hasta el punto de saber con exactitud que términos debe utilizar para optimizar su consulta.

Esto obliga a reconocer que, en la mayoría de las veces, la interpretación de la consulta formulada por el usuario requerirá algún tipo de inferencia adicional por parte del sistema. Estas inferencias deben dirigirse a contextualizar la consulta, pero este proceso hoy por hoy no puede ser completamente automático, porque en general un usuario no puede expresar sus objetivos de búsqueda, ya sea porque no tiene un objetivo preciso, ya sea porque su expresión es demasiado compleja como para ser interpretada por un sistema, ya sea porque la información se encuentra fragmentada en un conjunto de documentos.

Si bien no es posible interpretar completamente una consulta descontextualizada, sí resulta posible que el sistema realice los procesamientos lingüísticos más elementales y ofrezca sus resultados parciales al usuario como medio adicional y no sustitutivo de contextualizar e interpretar su consulta, es decir, de acceder a la información que busca.

La hipótesis que se utiliza en *Website Term Browser* es que el usuario busca un concepto o entidad, y como tal, su formulación lingüística se corresponde con un sintagma nominal. Este no es el caso en sistemas de búsqueda de respuestas (*Question Answering)*, en los que la formulación de la consulta es una proposición interrogativa que puede ir dirigida a localizar eventos, procesos, etc., pero es una hipótesis que encaja bastante bien en los sistemas de búsqueda basados en palabras clave.

Así, las palabras de la consulta se interpretan como las componentes de un sintagma que de forma ideal enuncia el objeto de la búsqueda. Evidentemente será muy difícil que ese sintagma aparezca en la colección con las mismas palabras que ha utilizado el usuario, y por ello el sistema deberá realizar las inferencias oportunas para encontrar las expresiones correspondientes tal como aparecen en la colección, incluso en el ejemplo extremo de que se encuentre en un idioma diferente.

Se dejarán fuera de este trabajo otras inferencias de nivel semántico y pragmático que, sin embargo, son interesantes para enriquecer el proceso de interacción con el usuario, como por ejemplo, inferencias sobre conocimiento organizativo o inferencias sobre los supuestos que se han utilizado recientemente en la interacción.

El sistema desarrollado en este trabajo se dirige a proporcionar al usuario un nivel parcial de procesamiento lingüístico resultado de un proceso inferencial capaz de recuperar variaciones morfosintácticas, semánticas y translingües de la consulta. Este nuevo nivel de acceso a la información se organiza en una nueva área adicional al tradicional listado de documentos. Ambas áreas (términos y documentos) se organizan de forma diferente según el prototipo y su interfaz.

4.4.1 Área de términos

Es probable que la recuperación y ordenación de sintagmas proporcione alguno que no sea relevante para el usuario, como también es posible que haya documentos irrelevantes en un ranking de documentos tradicional. Sin embargo, discriminar sintagmas es mucho más rápido que discriminar documentos.

Por esta razón, al usuario se le ofrece la posibilidad de seleccionar los sintagmas verdaderamente relevantes sin importar que alguno de los sintagmas que se le ofrecen no sea relevante o, incluso, que sea incorrecto: de forma rápida los detectará e ignorará. Esa es la virtud del sintagma nominal como expresión suficientemente concreta de un concepto, su poder informativo y su escasa ambigüedad permiten una discriminación sencilla y efectiva.

La selección de sintagmas relevantes puede verse imposibilitada si al usuario se le ofrecen un número excesivo de sintagmas sin orden ni estructura alguna y, por ello, es necesario que el sistema realice una preselección y organización según algún criterio de relevancia. Los criterios utilizados de relevancia, ordenación y selección de sintagmas ya se han mencionado en los apartados anteriores. De cara a la interacción, estos criterios se verán reflejados en las distintas interfaces.

Las acciones que puede realizar el usuario sobre el área de términos son sencillas y directas, requisito fundamental para que los usuarios acepten su uso: seleccionar un grupo para explorar los sintagmas que contiene, seleccionar un término para explorar los documentos que lo contienen y seleccionar un sintagma para formular una nueva consulta al buscador.

4.4.2 Área de documentos

El objetivo de la propuesta no es ofrecer una alternativa al ranking de documentos, sino complementarlo permitiendo al usuario varias vías de acceso a la información. Los sintagmas seleccionados y mostrados al usuario se convierten en buenos

términos de redefinición de la consulta, su mera selección permite incorporar información con el fin de afinar el ranking de documentos.

De esta forma, al usuario se le ofrece una segunda área de resultados, una lista ordenada de documentos relevantes para su consulta. De esta forma, tras la consulta, el usuario puede interactuar con las opciones que le ofrece tanto el área de términos como el área de documentos.

Capítulo 5

Website Term Browser

El modelo propuesto en el capítulo anterior se ha llevado a la práctica mediante el sistema *Website Term Browser* (WTB). Este sistema se ha construido de forma incremental mediante el desarrollo de una serie de prototipos sobre distintas colecciones. A continuación se describe la metodología de desarrollo (5.1), el trabajo de desarrollo preliminar de Extracción Automática de Terminología a partir del cual arranca el desarrollo del primer prototipo (5.2) y el proceso de implementación de los cinco prototipos del sistema (5.3), (5.4), (5.5), (5.6), (5.7). Aunque la evolución de los prototipos implica su evaluación para poder definir el siguiente refinamiento en el proceso de desarrollo, se ha dejado para el siguiente capítulo todas las cuestiones relacionadas con la evaluación final del sistema.

5.1 Metodología de desarrollo

Cuando resulta difícil tener claros todos los requisitos del sistema al inicio del desarrollo, es necesario facilitar la detección de errores de concepción y el descubrimiento de nuevos requisitos inadvertidos en las fases de análisis. Este es el caso de los desarrollos con carácter de investigación y, en concreto, de este trabajo. Teniendo esto en cuenta, el desarrollo del sistema ha seguido un ciclo de vida basado en la construcción de prototipos. Este modelo se hace apropiado, además, por tratarse del desarrollo de una aplicación que debe considerar la interacción con los usuarios.

El número inicial de prototipos viene determinado por el objetivo de implementar un buscador que considere variaciones morfosintácticas, semánticas y translingües de la consulta, y capaz de tratar con un gran volumen de información textual. Se debe considerar no sólo incertidumbre en las especificaciones finales del sistema, sino también respecto a los problemas técnicos al manejar grandes cantidades de información.

Los prototipos necesarios para el desarrollo completo del sistema son:

1. Prototipo monolingüe, sobre una colección pequeña (1.000 documentos), en el que se estudia la extracción de sintagmas y su recuperación pero sin realizar inferencias adicionales que lleven a la consideración de variaciones terminológicas.
2. Prototipo monolingüe, añadiendo expansión por sinónimos de los lemas de la consulta para considerar variaciones semánticas.
3. Prototipo multilingüe manejando tres idiomas, sobre una colección mayor en un orden de magnitud (de 10.000 a 20.000 documentos), donde además de expansión por sinónimos se realice también traducción de los lemas de la consulta.
4. Prototipo multilingüe manejando cinco idiomas, sobre una colección mayor en un orden de magnitud (unos 100.000 documentos) con toda la funcionalidad y estudiando la forma de organizar la terminología por idiomas.
5. Prototipo multilingüe sobre un entorno real de trabajo, con usuarios reales en el dominio UNED.es (40.000 documentos aproximadamente), registrando la interacción de los usuarios y permitiendo obtener datos cuantitativos para evaluar el sistema.

5.1.1 Colecciones de prueba

Los prototipos del sistema se han aplicado a colecciones que cumplían las especificaciones de cada uno de ellos.

5.1.1.1 Colección monolingüe de recursos educativos

La *colección monolingüe de recursos educativos* es una colección dirigida a la extracción de terminología en el dominio concreto de educación primaria y secundaria. Esta colección contiene únicamente 1075 documentos con un total de 670.646 palabras. Principalmente se trata de páginas web que describen recursos educativos multimedia. Es una colección pequeña y bastante homogénea que se ha

construido mediante un crawler que ha recopilado las páginas de dos sitios web relacionados con el dominio:

- Programa de Nuevas Tecnologías:
  http://www.pntic.mec.es/main_recursos.html

- Aldea Global:
  http://sauce.pntic.mec.es/~alglobal

A ello se le ha añadido un documento extenso en formato electrónico elaborado por el Instituto Cervantes y que pertenece al dominio en estudio.

5.1.1.2 Colección multilingüe de noticias internacionales

La *colección multilingüe de noticias internacionales* tiene 12.156 documentos y unos 43 Mb de texto. Se compone de tres fuentes distintas en español, inglés y catalán:

- Noticias en español: proceden del periódico El País en formato electrónico y accesible a través de Internet. Consta de 7.364 documentos y 23,7 Mb.
- Noticias en inglés: proceden de la versión electrónica del Washington Post también accesible desde Internet. Consta de 3.050 documentos y 16,2 Mb.
- Noticias en catalán: proceden de la versión electrónica de El Periódico accesible desde web. Consta de 1.742 documentos y 3,1 Mb.

Estas noticias fueron recopiladas entre abril de 1998 y mayo de 1999.

5.1.1.3 Colección multilingüe de recursos educativos en Europa

La colección multilingüe de recursos educativos se ha construido a partir de las páginas web de diversos repositorios en toda Europa. Esta colección consta de 92.172 documentos en español, inglés, alemán, francés, italiano y catalán, número cercano a los 100.000 documentos especificados en el cuarto prototipo, para el que se destina la colección. Sin embargo, por no disponer de herramientas lingüísticas apropiadas para el alemán, los 26.617 documentos correspondientes a este idioma no se han indexado en ningún prototipo. La distribución de los restantes 65.555 documentos (aprox. 200 Mb) es la siguiente:

- Español, 6.271 documentos
- Inglés, 12.631 documentos
- Francés, 12.534 documentos
- Italiano, 10.970 documentos
- Catalán, 23.149 documentos.

El número final de documentos de esta colección (65.555) es mucho menor que los 100.000 documentos especificados para el cuarto prototipo. Esto ha llevado al planteamiento de un trabajo adicional para evaluar la escalabilidad del sistema. Este trabajo se presenta en el capítulo de evaluación (**¡Error! No se encuentra el origen de la referencia.**).

5.1.1.4 Colección de páginas web en el dominio uned.es

La colección de páginas web en el dominio UNED.es ocupa 845 Mb de código HTML que se convierten en 153 Mb de texto a considerar en la indexación. Esta colección tiene 41.949 documentos diferentes una vez desechadas las páginas repetidas y las que no contienen información textual.

5.1.2 Elección de la arquitectura y entorno tecnológico

El volumen de la información que se maneja, así como la dificultad de migrar los recursos y herramientas lingüísticos obliga a una arquitectura distribuida en la que el usuario pueda interactuar con el sistema de forma remota. El modelo cliente-servidor se ajusta a las necesidades de interacción siendo suficiente el protocolo *http*. De esta forma, el usuario puede utilizar cualquier navegador de Internet para comunicarse con el sistema, sin necesidad de instalar ningún componente adicional en su ordenador.

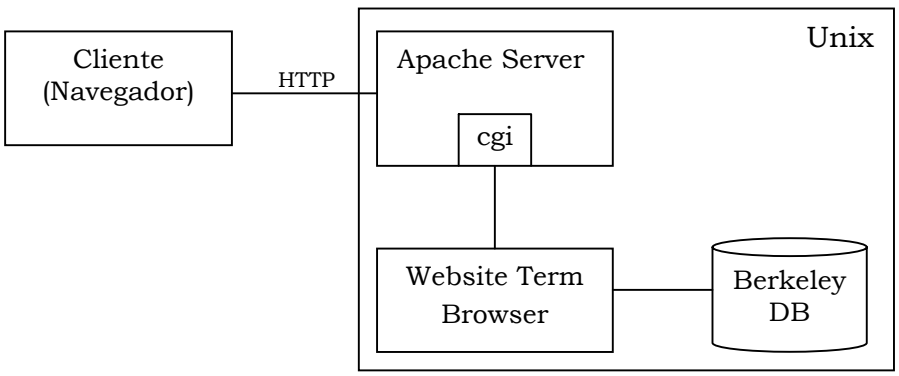


Figura 5-1. Entorno tecnológico de WTB

5.1.3 Determinación del contexto y alcance del sistema

El sistema debe recuperar de una base de documentos textuales en diversos idiomas, aquellos textos relacionados con una consulta con independencia del idioma de la consulta y de los documentos. Es decir, una consulta en un idioma debe proporcionar acceso a los documentos relevantes en otro idioma.

Además, cuando los prototipos puedan utilizarse en un entorno real de trabajo, el sistema debe registrar su interacción con los usuarios con el fin de obtener datos que permitan la evaluación no sólo cualitativa sino también cuantitativa de la utilidad del sitema.

La construcción del sistema lleva consigo necesariamente la utilización de programas para recolectar las colecciones de documentos y páginas web (crawler). No es objetivo del trabajo desarrollar nuevas aproximaciones en este sentido, por lo que se utilizarán aplicaciones ya existentes como *wget*.

El contexto del sistema viene determinado por el siguiente diagrama de contexto:

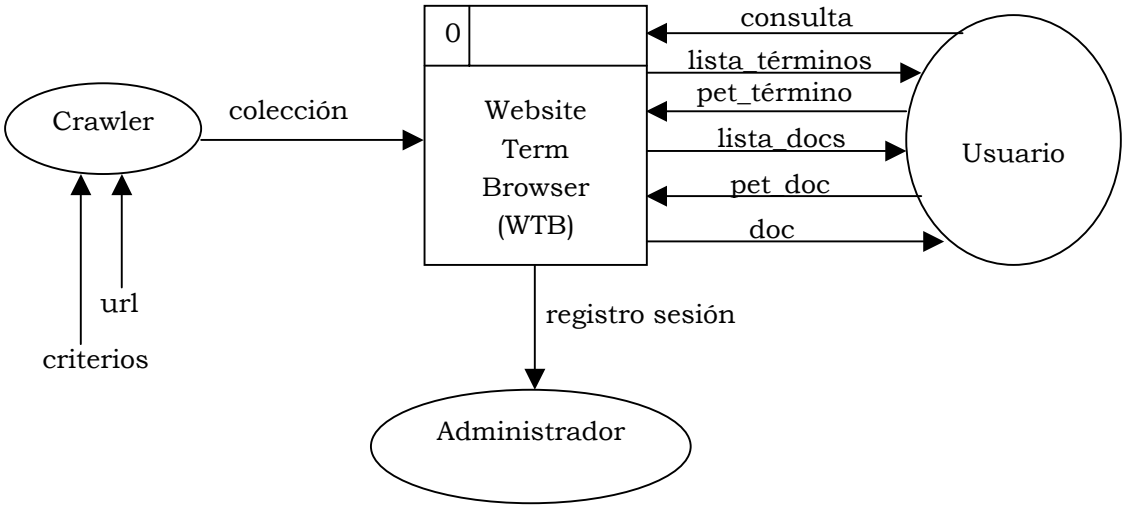


Figura 5-2. Diagrama de contexto de WTB.

5.1.4 Modelo lógico de datos

Los modelos propuestos de indexación, recuperación e interacción determinan el modelo lógico de datos. Las relaciones entre palabras, lemas sintagmas, documentos y colecciones vienen definidas por el siguiente diagrama:

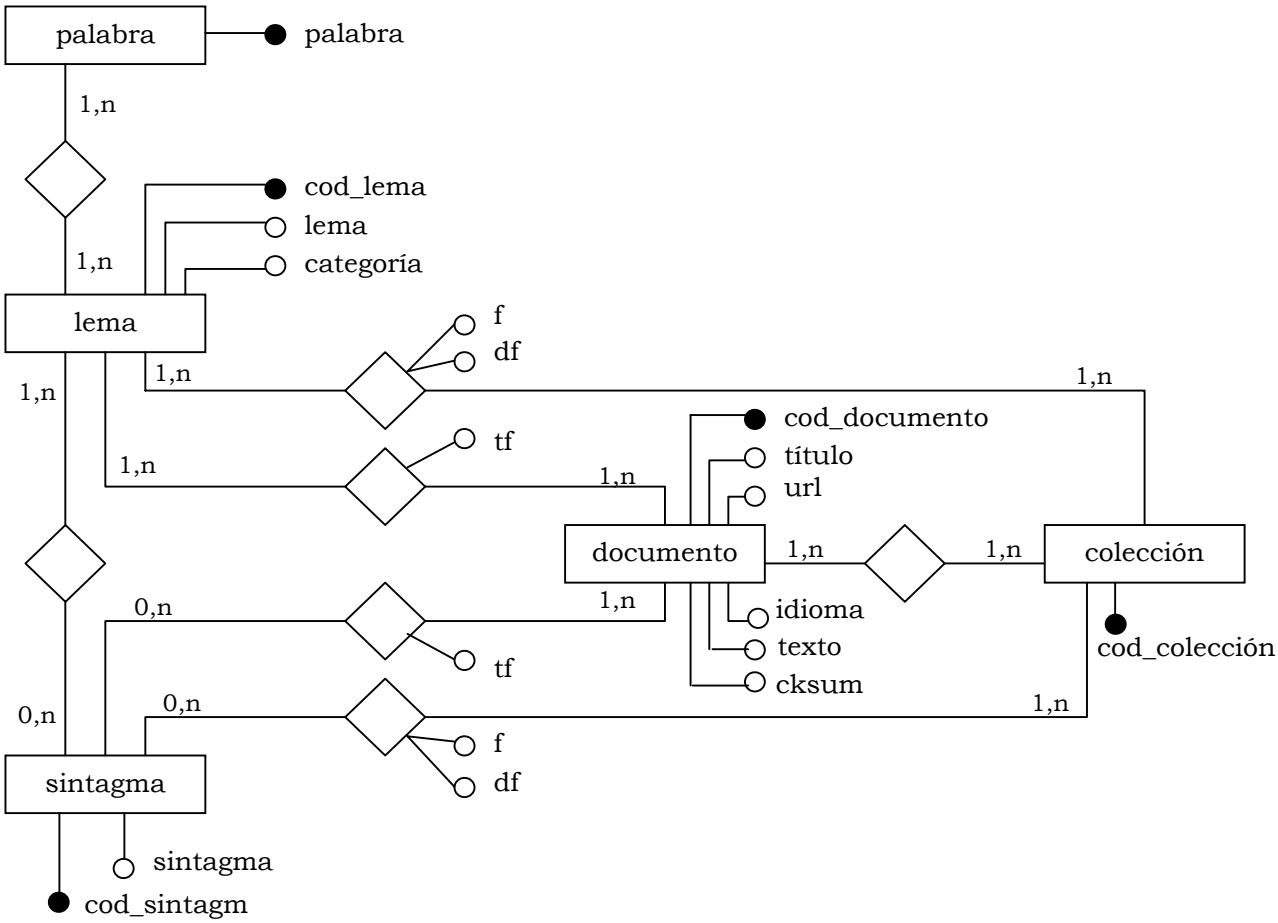


Figura 5-3. Modelo lógico de datos.

donde

título: es el título del documento,
url: es la dirección donde se ubica el documento original,
idioma: es el idioma del documento,
texto: es el contenido textual del documento una vez filtrado,
cksum: valor numérico que permite identificar documentos repetidos,
f: es la frecuencia de un término en una colección,
df: es el número de documentos de la colección que contienen al término,
tf: es la frecuencia de aparición de un término en un documento.

Las relaciones entre las entidades del modelo son las siguientes:

- A cada palabra le corresponde al menos un lema y a cada lema le corresponde al menos una palabra.
- Cada sintagma contiene al menos un lema pero un lema puede que no participe en ningún sintagma.
- Los lemas están presentes en al menos un documento y los documentos al menos contienen un lema. Lo mismo con los lemas respecto a las colecciones.
- Un sintagma está presente en al menos un documento, pero un documento puede no contener ningún sintagma. Lo mismo ocurre con los sintagmas respecto a las colecciones.
- Un documento aparece en al menos una colección y una colección tiene al menos un documento.

5.1.5 Comportamiento dinámico de la interfaz de usuario

En todos los prototipos, la interfaz de usuario tiene tres áreas. En la primera, el usuario introduce la consulta en forma de palabras clave y solicita la búsqueda. En la segunda, el sistema le ofrece al usuario una lista de documentos relacionados con la consulta. Desde esta lista se puede acceder al contenido de los documentos. La tercera área es una lista o jerarquía de términos (palabras y sintagmas) relacionados con la consulta. Los usuarios pueden seleccionar estos términos o bien para construir una nueva lista de documentos o para utilizarlo como nueva consulta. Los detalles concretos de interacción dependen del prototipo y se muestran más adelante.

5.2 Extracción Automática de Terminología

El sistema de recuperación translingüe propuesto se basa en la utilización de sintagmas como elementos de indexación de la colección y como medio para reducir la ambigüedad de traducciones y expansiones de la consulta. Como los sintagmas deben extraerse de forma automática a partir de las colecciones, el desarrollo preliminar del sistema se ha dirigido a estudiar la metodología de Extracción Automática de Terminología (TE). Esta metodología conduce de forma natural a considerar el grado de lexicalización de los sintagmas como uno de los criterios más importantes para su selección y posterior organización en las interfaces.

Con este trabajo se pretenden desarrollar las técnicas utilizables posteriormente en la indexación IR. Sin embargo, existen diferencias entre las tareas y objetivos de TE

e IR que serán necesarios estudiar con el fin de adaptar y desarrollar las técnicas finales que se emplearán en los prototipos del sistema WTB.

Este trabajo de TE fue una de las tareas desarrolladas dentro del proyecto europeo European Schools Treasury Browser[10] (ETB), con el objetivo de extraer la lista de términos españoles candidatos a formar parte del tesauro que permite organizar y recuperar los recursos en el dominio de educación primaria y secundaria, y el ámbito de las nuevas tecnologías. Si bien ya existían tesauros multilingües para el contexto de la educación, el objetivo de las primeras fases del proyecto ETB fue enriquecerlos con términos nuevos emergentes con la incorporación de las nuevas tecnologías en el ámbito educativo. Además, el dominio de las nuevas tecnologías en la educación se trata de un núcleo adecuado para un tesauro multilingüe en el que deben recogerse las diferentes culturas educativas europeas.

El trabajo permitió desarrollar métodos y técnicas generales para obtener listas terminológicas independientemente del dominio de aplicación, y ser utilizadas en recuperación de información multilingüe. La tarea de extracción automática de listas terminológicas tiene como objetivo identificar y extraer expresiones, tanto de una como de varias unidades léxicas, que se utilizan de forma habitual como referentes de los conceptos del dominio. Si el dominio del problema es bastante específico, como es el caso, en general no se dispondrán de listas terminológicas ya elaboradas. Tampoco será suficiente utilizar tesauros y recursos de dominios generales como EuroWordNet.

La metodología de extracción automática de terminología sigue tres pasos:

1. *Detección de términos* candidatos de un corpus (en este caso la colección de documentos) mediante procesamiento del lenguaje natural (PLN): análisis morfosintáctico, etiquetado de categorías gramaticales y análisis sintáctico superficial.

2. *Pesado de términos* mediante alguna medida de relevancia a partir de los datos estadísticos obtenidos de la colección.

3. *Selección de términos* tras la ordenación de los términos según su relevancia, eliminando los términos que no se ajusten a unos umbrales de relevancia determinados.

A estas fases, hay que añadirles una fase previa en la que se prepara la colección sobre la que se extrae la terminología.

[10] http://www.eun.org/etb

Debido a que su tratamiento será diferente, se va a distinguir entre la extracción de términos de una sola palabra (monoléxicos) y la extracción de sintagmas o términos compuestos por varias palabras (poliléxicos).

En los siguientes apartados se explica el desarrollo del proceso en las fases mencionadas.

5.2.1 Preparación de las colecciones

La tarea de Extracción de Terminología consiste en identificar los términos que se utilizan habitualmente en un dominio específico para referirse a los conceptos propios de dicho dominio. Para realizar esta tarea, se han aplicado técnicas automáticas que requieren la previa recopilación de una colección de documentos pertenecientes al dominio de trabajo. La colección utilizada es la *Colección Monolingüe de Recursos Educativos*.

Asimismo, con el fin de no considerar términos que aparezcan frecuentemente en otros dominios y, por tanto, no sean específicos del dominio en estudio, se ha utilizado una segunda colección compuesta por las páginas web de noticias internacionales del periódico El País (los documentos en español de la *Colección Multilingüe de Noticias Internacionales*). Esta última colección permite detectar:

- Palabras que por ser demasiado frecuentes en todos los dominios no son relevantes para ninguno. Principalmente son lo que se denominan "stopwords" y muchas de ellas pertenecen a categorías gramaticales cerradas.

- Palabras propias del dominio. Es decir, las palabras más características del dominio de educación secundaria no deberían aparecer con frecuencia en una colección genérica o de otro dominio como el de noticias internacionales.

Ya se ha comentado que ambas colecciones se componen principalmente de páginas en formato *HTML* recogidas mediante un *crawler*. Las colecciones que se obtienen no pueden utilizarse sin un procesamiento previo que elimine elementos que distorsionan el posterior proceso de Extracción Automática de Terminología. A continuación se detalla el procesamiento previo tanto para la colección de noticias internacionales como la de recursos educativos.

5.2.1.1 Eliminación de etiquetas HTML

En primer lugar, debe eliminarse el código html que contienen las páginas. Etiquetas y atributos HTML constituyen cadenas que no deben considerarse en la identificación de términos relevantes. Para ello se ha utilizado el comando *unhtml* con licencia GNU. Para comprobar el correcto funcionamiento del comando se han explorado los documentos procesados con el mismo, comprobándose que su funcionamiento es adecuado para casi la totalidad de las páginas. El mayor problema en la conversión de documentos html a texto es que en ocasiones la estructura del documento no siempre se corresponde con la visualización. Así, aunque visualmente parezca un texto coherente, no es así en la estructura interna del documento. Esto puede ocurrir, por ejemplo, con el uso de tablas.

5.2.1.2 Eliminación de páginas en otros idiomas

Algunas páginas no están escritas en español (por ejemplo, porque son un recurso de enseñanza de lenguas extranjeras). Estas páginas no sólo no van a aportar terminología en español relevante para el dominio, sino que pueden producir terminología incorrecta. Por tanto, estas páginas deben eliminarse.

Para identificar el idioma de un texto, puede ser suficiente con utilizar la lista de palabras más frecuentes de dicho idioma. Si en un texto aparecen en una proporción suficiente, el texto queda identificado. Sin embargo, esta técnica no funciona con suficiente precisión si se trata de documentos pequeños como ocurre con muchas de las páginas web.

Por esta razón, se ha utilizado un detector de idioma basado en n-gramas. Esta herramienta toma un texto y devuelve los idiomas más probables en los que está escrito. Como los sitios web son españoles, la mayoría de las páginas están en español. Por esta razón, basta con que el detector de idioma considere que uno de los idiomas posibles del texto sea el español como para aceptar el documento. El resto se desechan.

La evaluación de una muestra indicó que la precisión de la herramienta es adecuada, desechándose documentos no escritos en español y documentos demasiados pequeños e incoherentes (cabeceras, menús, etc.) que no aportan información relevante para la extracción de terminología.

5.2.1.3 Eliminación de páginas repetidas

Debido a que los sitios web suelen estar en continua revisión de sus contenidos, es corriente que aparezcan páginas con diferentes nombres pero con exactamente el mismo contenido, es decir, páginas que deben considerarse repetidas.

La aparición de secuencias idénticas de palabras en lugares diferentes indica la posibilidad de que se traten de expresiones terminológicas. Sin embargo, en el caso de páginas repetidas no es así, es decir, se produce una distorsión en la extracción de terminología. Por tanto, para la tarea de extracción de terminología deben eliminarse los documentos repetidos.

La coincidencia en la longitud de un fichero no es evidencia suficiente de que un documento esté repetido, sin embargo sí lo es junto al valor de CRC (código de redundancia cíclica) asociado al fichero devuelto por una función de *checksum.* Se ha utilizado el comando *chksum* de UNIX con muy buenos resultados.

5.2.2 Detección de términos

5.2.2.1 Detección de términos monoléxicos

La detección de términos monoléxicos se realiza sobre la base de su categoría gramatical. Sólo se consideran nombres, adjetivos y verbos. Para ello, es necesario etiquetar los textos morfosintácticamente.

Sin embargo, antes es necesario ignorar abreviaturas, palabras de otros idiomas y cadenas de caracteres que no constituyen palabras. Para ello se han utilizado patrones sencillos y los diccionarios en formato electrónico.

Para el etiquetado morfosintáctico de los textos se ha utilizado el tokenizador, el analizador morfológico y el etiquetador morfosintáctico ya descritos con anterioridad (*MACO+, Relax*).

Tras la fase de detección de términos monoléxicos se obtiene una lista de lemas con el número de ocurrencias de cada uno de ellos en la colección (en cualquiera de sus formas). Además, a cada lema se le asocia el número de documentos diferentes en los que aparece, así como la lista de los mismos.

5.2.2.2 Detección de términos poli- léxicos

Para detectar sintagmas nominales en un texto, es común realizar un análisis sintáctico superficial. El problema de extracción de terminología no requiere un análisis completo de las oraciones, basta con detectar las secuencias de palabras que pueden constituir un sintagma nominal. Algunos sintagmas nominales complejos pueden descomponerse a su vez en otros sintagmas nominales más simples. Para la extracción de terminología conviene considerar tanto unos como otros. Será en fases posteriores a la detección cuando se decida cuáles de ellos suponen lexicalizaciones propias del dominio. Por ejemplo, "profesor titular de educación especial" es un sintagma nominal compuesto a su vez por otros sintagmas nominales como son "profesor titular", "titular de educación", "educación especial", etc. En esta fase conviene extraer todos los sintagmas candidatos, sin importar la detección de sintagmas que puedan no ser adecuados desde un punto de vista semántico o terminológico. Será en fases subsiguientes cuando se desechen las expresiones que no suponen lexicalizaciones de los conceptos del dominio.

El hecho de detectar todas las secuencias de palabras que puedan constituir un sintagma nominal, además de conveniente, simplifica mucho la tarea, ya que se convierte en un problema de ajuste de patrones y no de segmentación o análisis sintáctico completo, procesos que requieren herramientas especializadas, dependientes del idioma, difíciles de encontrar y con un coste computacional alto.

Para detectar todos los posibles sintagmas nominales contenidos en una oración el uso de patrones sintácticos resulta adecuado. Los patrones permiten encontrar todos los sintagmas que se ajustan a ellos en todos los documentos de la colección. Para llevar a cabo esta tarea se ha implementado un reconocedor de patrones que utiliza las etiquetas morfosintácticas de las palabras dadas por el etiquetador.

El reconocedor de patrones toma como entrada, además del conjunto de patrones, un texto etiquetado morfosintácticamente según la cascada de herramientas ya mencionada en el caso de detección de términos monoléxicos: tokenizador, analizador, etiquetador. Los patrones se definen como secuencias de etiquetas morfosintácticas. Si en el texto aparece una secuencia de palabras cuyas etiquetas se ajustan a las del patrón, entonces se ha reconocido un nuevo candidato.

Los patrones utilizados no intentan cubrir todas las construcciones posibles de un sintagma nominal, sino sólo aquellas más comunes en expresiones terminológicas. Los patrones utilizados son los siguientes:

1. N N
2. N A
3. N PP
4. N [A] Prep N [A]
5. N [A] Prep Art N [A]
6. N [A] Prep V [N [A]]

en los que A es adjetivo, N es nombre, PP es participio, Prep es preposición, Art es artículo y V es verbo en infinitivo.

A la vez que se reconocen los patrones en los textos, se construye un índice que registra para cada compuesto el número de veces que aparece en la colección, el número de documentos diferentes en que aparece, así como la lista de documentos en que aparece.

5.2.2.3 Eliminación de términos poli-léxicos

Los patrones reconocen 72.453 secuencias de más de una palabra que conforman sintagmas nominales en la *Colección Monolingüe de Recursos Educativos*. Aproximadamente el 75% de estos sintagmas aparece únicamente una vez en toda la colección y la gran mayoría de ellos suponen expresiones que no son relevantes para el dominio o son secuencias erróneas.

Para las tareas de Recuperación de Información no es necesario desechar expresiones incorrectas pero, sin embargo, para extraer los términos específicos de un dominio es preferible que las expresiones identificadas sean correctas aún a costa de no identificar algunas expresiones. En otras palabras, es más importante mantener una *precisión* alta que un *nivel de recuperación* elevado.

Por esta razón, para la tarea de Extracción de Terminología, tras un proceso de inspección visual, se ha establecido un umbral en el número de ocurrencias de una expresión multipalabra: todas aquellas expresiones que aparezcan una sola vez son desechadas de antemano.

5.2.2.4 Eliminación de términos monoléxicos

La fase de pesado de términos establece un ranking de los mismos según alguna medida de relevancia. El acierto en los términos que finalmente se escojan o se desechen vendrá determinado por la bondad de la medida de relevancia utilizada.

Por esta razón, es conveniente utilizar criterios adicionales que pueden aplicarse de forma previa para realizar un primer sesgo de las listas de términos candidatos.

Los tres criterios utilizados son:

1. Eliminación de los términos muy poco frecuentes (UMBRAL_CORPUS=10). Los términos poco frecuentes en la colección tienen baja probabilidad de ser característicos del dominio.

2. Eliminación de los términos muy frecuentes en la colección de dominio no restringido. (UMBRAL_PAIS=1000). Los términos que aparecen con mucha frecuencia en otros dominios también tienen una probabilidad baja de ser representativos del dominio.

3. Eliminación de los términos que no participan en compuestos lexicalizados. La mayoría de los términos específicos de un dominio van a participar en compuestos multipalabra, ya que es la forma de restringir el significado del término.

Además, se apartan para un estudio posterior los términos que no aparecen en el diccionario monolingüe español. En su mayoría son cadenas incorrectas y son muy pocas las palabras interesantes, en su mayoría palabras técnicas y de dominio específico cuya detección sería posible mediante el uso de diccionarios técnicos.

Los umbrales de frecuencia para sesgar la lista de términos dependen del número de términos que se quieren contemplar en las subsiguientes fases. En este caso, se ha querido realizar un estudio manual de los candidatos para evaluar los métodos automáticos, de forma que los umbrales se ajustaron para obtener entre 2000 y 3000 términos.

5.2.3 Pesado de términos

5.2.3.1 Pesado de términos monoléxicos

El resultado de la fase anterior es la obtención de una lista de lemas candidatos con el número de ocurrencias asociado en la colección, y el número de documentos diferentes en los que aparece. El objetivo de la fase de pesado es asignar un peso de relevancia a cada uno de ellos para finalmente seleccionar únicamente los términos más relevantes para el dominio.

La medida de relevancia utilizada tiene en cuenta:

1. Frecuencia del término en la colección.
2. Número de documentos en los que aparece.
3. Frecuencia del término en un dominio genérico.

La medida que se propone para relacionar estas frecuencias es la siguiente:

$$\text{Relevance}(t, sc, gc) = 1 - \frac{1}{\log_2 \left(2 + \frac{F_{t,sc} \cdot D_{t,sc}}{F_{t,gc}} \right)}$$

Donde

$F_{t,sc}$: frecuencia relativa del término *t* en la colección específica *sc* (número de ocurrencias de *t* en *sc* dividido por el número de palabras de *sc*).

$F_{t,gc}$: frecuencia relativa del término *t* en la colección genérica *gc* (número de ocurrencias de *t* en *gc* dividido por el número de palabras de *gc*).

$D_{t,sc}$: número relativo de documentos de la colección específica *sc* donde aparece el término *t* (número de documentos de *sc* donde aparece *t* dividido por el número de documentos de *sc*).

La medida otorga mayor peso a aquellos términos frecuentes en la colección específica, infrecuentes en la colección genérica, premiando a aquellos términos que aparecen en varios documentos diferentes de la colección específica.

Los términos que son frecuentes en la colección del dominio tienen una probabilidad alta de ser conceptos lexicalizados del dominio. Sin embargo, muchos de estos términos pueden ser muy frecuentes también en otros dominios y, por tanto, no deberían considerarse como característicos del dominio en estudio. Si son muy frecuentes en todos los dominios también lo serán en la colección genérica y, llegado este punto, ya habrán sido eliminados. Pero si se trata de términos de un dominio específico que no es el de estudio y que aparece con frecuencia en algún documento (e.g. un texto educativo sobre mitología) se corre el peligro de ser aceptado como término propio del dominio en estudio. Para solucionar este problema resulta muy efectivo considerar en la medida de relevancia el número de documentos en los que aparece el término. Un término relevante para el dominio

debe aparecer en varios documentos del dominio. Cuantos más documentos del dominio contengan a ese término más relevante para el dominio será el término.

En resumen, la medida considera los siguientes casos:

- Términos poco frecuentes en la colección del dominio: se eliminan.

- Términos muy frecuentes en la colección del dominio:

    - Muy frecuentes en todos los dominios (stopwords): se eliminan
    - Muy frecuentes en muy pocos documentos del dominio: con probabilidad alta pertenecen a otro dominio: se eliminan.
    - Frecuentes en varios documentos de la colección del dominio: con probabilidad alta son términos característicos del dominio.

Como se observa, la medida de relevancia tiene en cuenta estos criterios para ordenar los lemas en un ranking de mayor a menor relevancia.

5.2.3.2 Pesado de términos poli-léxicos

La medida anterior no se puede utilizar para el caso de los términos de varias componentes porque las frecuencias de aparición en las colecciones son muy pequeñas. La consideración de los valores de relevancia de las componentes del compuesto poliléxico será el aspecto determinante en el momento de la selección de los términos.

5.2.4 Selección de términos

En el momento de seleccionar la lista de términos finalmente aceptados es necesario, una vez más, distinguir entre términos mono y poliléxicos. Esto es debido a que las medidas de relevancia son diferentes en los dos casos.

5.2.4.1 Selección de términos monoléxicos

La fase anterior asigna un peso de relevancia a cada uno de los términos candidatos. Ordenando esta lista según este peso se puede establecer un cierto umbral que permita seleccionar aquellos términos más relevantes. El valor del umbral depende del número de términos que quieran seleccionarse finalmente.

5.2.4.2 Selección de compuestos léxicos

La selección de los compuestos léxicos relevantes se ha realizado atendiendo a la relevancia de sus componentes. Mediante la exploración manual de una muestra se comprobó que aquellos compuestos en los que ninguna de las componentes pertenecía a la lista de términos monoléxicos seleccionados eran compuestos poco o nada relevantes.

De esta forma, se han seleccionado aquellos compuestos poliléxicos para los cuales alguna de sus componentes había sido considerada relevante.

5.2.5 Evaluación

5.2.5.1 Entorno de exploración visual de los resultados

La exploración visual de los resultados es necesaria a lo largo de todo el proceso:

1. Para ayudar a tomar decisiones de desarrollo y refinamiento del proceso.
2. Para evaluar la bondad de los métodos, medidas y técnicas utilizadas, y sugerir modificaciones y mejoras.
3. Como herramienta para el análisis y filtrado por parte de los documentalistas.

El uso de documentos html para mostrar los resultados del proceso permite una exploración sencilla, intuitiva y ordenada gracias al uso de hiperenlaces. El inconveniente es que la generación de las páginas puede ser bastante costosa y no se justifica si deben generarse manualmente para cada resultado intermedio.

Sin embargo, para este trabajo las páginas HTML que permiten explorar los resultados parciales se generan automáticamente en cada iteración del proceso. La página HTML de la *Figura 5-4* muestra los datos que permiten calcular la medida de relevancia de cada uno de los términos mono-léxicos (frecuencia en la colección del dominio, número de documentos en los que aparece, frecuencia en la colección de noticias, número de compuestos en los que interviene, peso y, finalmente, si está presente o no en un diccionario). Además, cada término monoléxico incluye un hiperenlace a los contextos (kwic) en los que aparece y otro a la lista de compuestos (compounds) léxicos en los que participa en cualquiera de sus variantes morfosintácticas.

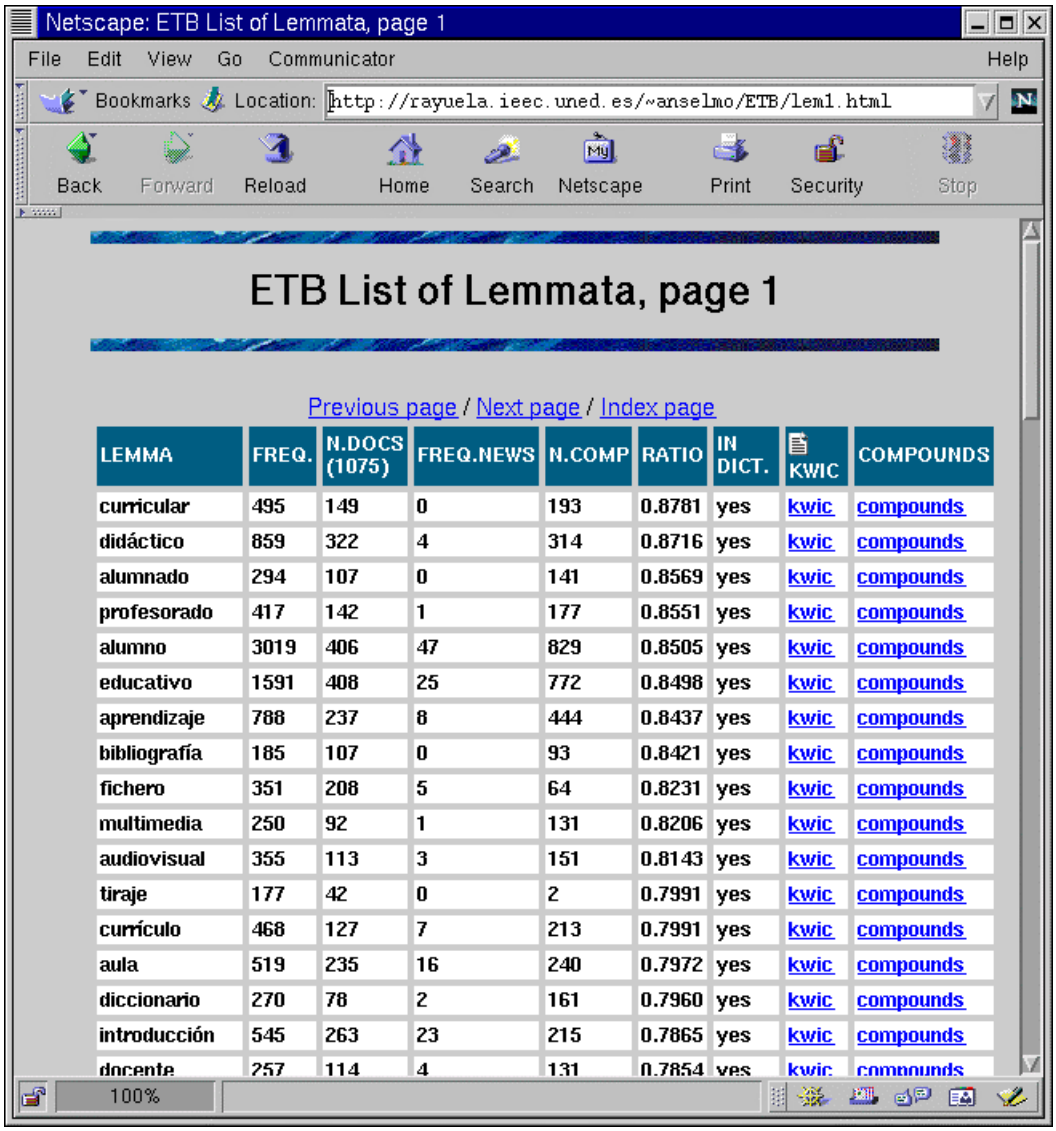
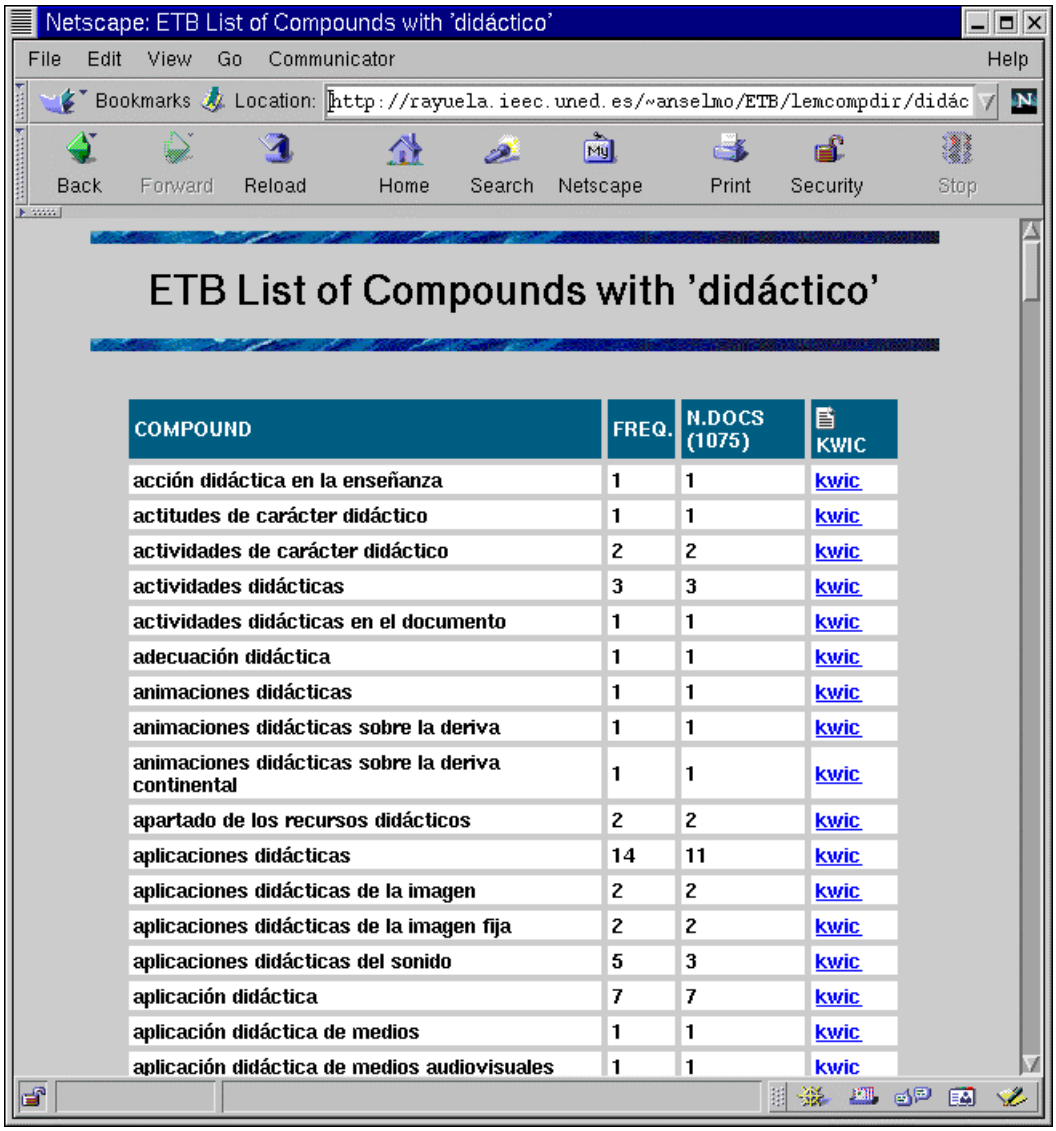


Figura 5-4. Visualización las listas terminológicas (términos mono-léxicos)

La *Figura 5-5* muestra el tipo de páginas correspondientes a los compuestos léxicos, con la información de frecuencias y documentos en los que aparece, así como un hiperenlace a sus contextos en la colección. Obsérvese que se relaciona el término monoléxico con los compuestos léxicos en cualquiera de las formas variantes del término monoléxico.



COMPOUND	FREQ.	N.DOCS (1075)	KWIC
acción didáctica en la enseñanza	1	1	kwic
actitudes de carácter didáctico	1	1	kwic
actividades de carácter didáctico	2	2	kwic
actividades didácticas	3	3	kwic
actividades didácticas en el documento	1	1	kwic
adecuación didáctica	1	1	kwic
animaciones didácticas	1	1	kwic
animaciones didácticas sobre la deriva	1	1	kwic
animaciones didácticas sobre la deriva continental	1	1	kwic
apartado de los recursos didácticos	2	2	kwic
aplicaciones didácticas	14	11	kwic
aplicaciones didácticas de la imagen	2	2	kwic
aplicaciones didácticas de la imagen fija	2	2	kwic
aplicaciones didácticas del sonido	5	3	kwic
aplicación didáctica	7	7	kwic
aplicación didáctica de medios	1	1	kwic
aplicación didáctica de medios audiovisuales	1	1	kwic

Figura 5-5. Visualización las listas terminológicas (términos poli-léxicos)

La *Figura 5-6* muestra los contextos de un término en la colección, permitiendo el acceso a los documentos ya procesados (versiones texto y texto etiquetado) para un análisis más profundo del contexto del término. De esta forma se pueden conocer los sentidos en los que se utiliza el término dentro de la colección, permitiendo relacionarlo con los términos correspondientes en los otros idiomas del tesauro multilingüe.

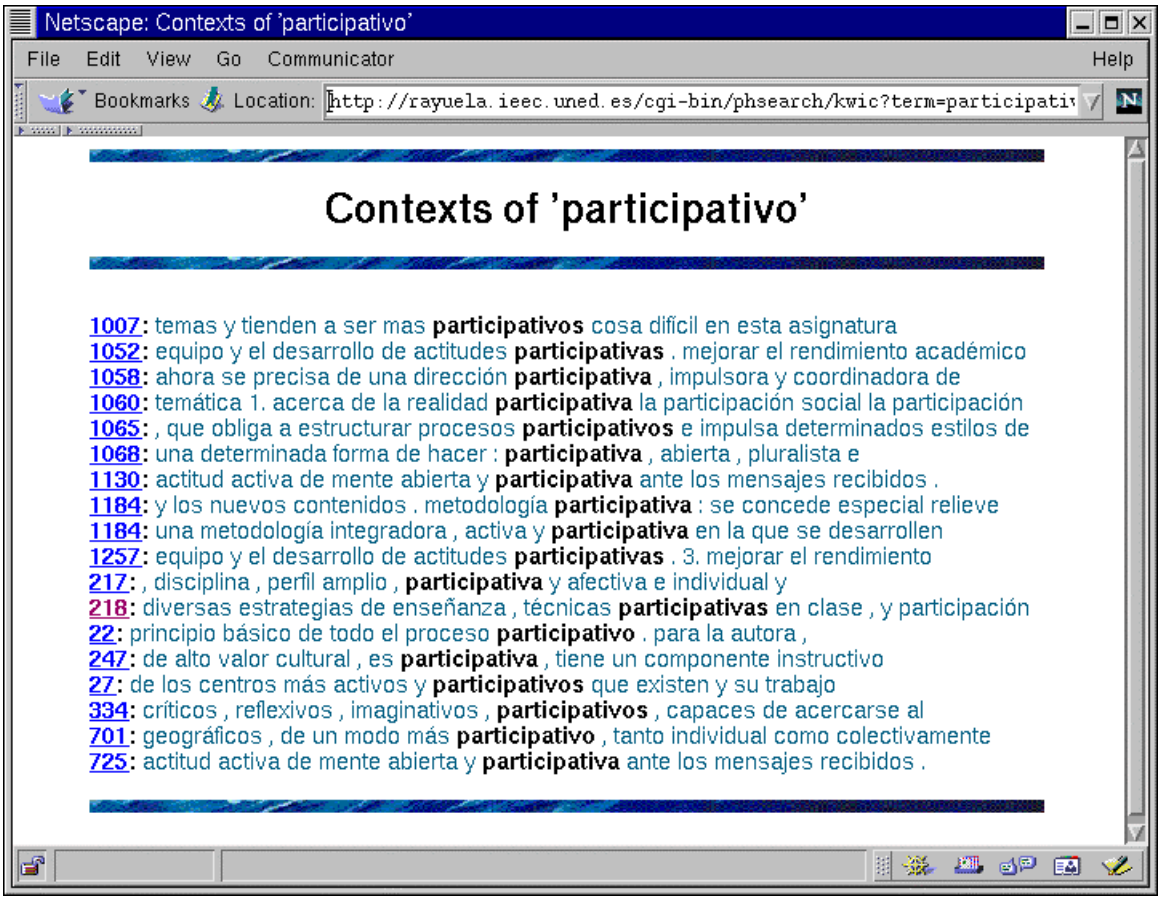


Figura 5-6. Visualización de los contextos de un término en la colección

5.2.5.2 Revisión interactiva de los términos seleccionados automáticamente

Para facilitar la clasificación de los términos se desarrolló un pequeño interfaz basado en el tratamiento de *XML* que permite el editor *EMACS*. Como candidatos a la lista final términos se seleccionaron 2856 términos que fueron revisados con ayuda de esta herramienta y clasificados según las siguientes clases:

1. **Término incorrecto**: el término obtenido tras el procesamiento automático no es una expresión aceptable en español. Por ejemplo, *profesorado materiales*.

2. **Término no lexicalizado**: es una expresión aceptable en español, pero no se trata de un uso comúnmente adoptado para designar un concepto. Por ejemplo, *alumnos ingleses*.

3. **Término fuera del dominio**: es una expresión aceptable y lexicalizada, pero no interesa para el dominio en estudio, en este caso el de *recursos educativos en primaria y secundaria*. Por ejemplo, *biblioteca nacional*.

4. **Término adecuado**: es una expresión aceptable, lexicalizada y relevante para el dominio de estudio. Por ejemplo, *proyecto curricular*.

5. **Término de dominio específico**: es un término adecuado pero considerado demasiado específico. Por ejemplo, *ciencias sociales* se ha considerado término específico.

6. **Término de dominio específico informático**: es un término de dominio específico relativo a la informática. Se ha decidido distinguir este dominio dada su relevancia para la terminología de recursos educativos en la web. Además, se trata de términos que seguramente no se contemplan en tesauros ya existentes relevantes para el dominio de estudio. Por ejemplo, *sistema operativo*.

7. **Variantes**: se trata de términos adecuados, pero con componentes flexionadas de forma diferente (plurales o singulares) de forma que existe otro término ya contemplado y que se prefiere a éste. Por ejemplo, *proyectos curriculares*.

Estas clases se derivan de las características del tesauro que se quiere construir.

Adecuado	Dominio específico	Informática	Variantes
1235 43%	513 18%	59 2%	78 3%

Tabla 5-1. Términos adecuados tras la extracción automática

Incorrectos	No lexicalizados	No dominio
151 5%	515 18%	305 11%

Tabla 5-2. Términos no adecuados tras la extracción automática

La *Tabla 5-1* y *Tabla 5-2* muestran como se han clasificado los 2856 términos candidatos. El resultado de la exploración manual ha permitido refinar los resultados de la extracción automática, desechando 971 términos (34%) no

adecuados y clasificando al resto según su grado de especificidad. Puede observarse que el 66% de los términos son aceptables, si bien resulta necesario hacer algún tipo de distinción sobre su adecuación al dominio y a su grado de especificidad.

Es destacable que estos porcentajes no son uniformes en toda la lista de términos candidatos. Si se ordenan por su valor de relevancia, el porcentaje de términos adecuados es mucho mayor en lo alto de la lista. Esto implica que la precisión total disminuye cuanto mayor sea el número de candidatos que se contemplen. Si bien puede parecer que un 66% de términos aceptables en distinto grado es un porcentaje bajo, hay que destacar que el número de candidatos que se han explorado manualmente (2856) es bastante elevado. Considerando un menor número de candidatos la precisión de los resultados se incrementa.

5.3 Primer Prototipo

El primer prototipo de WTB recupera documentos de la colección monolingüe de recursos educativos utilizada en los trabajos previos de extracción de terminología. En este prototipo se exploró la posibilidad de aprovechar con fines de acceso a la información la terminología extraída de forma automática a partir esta colección. Esto implica que han sido etiquetados los textos completos con la categoría de las palabras y que los patrones y el proceso de extracción de sintagmas son los mismos que en el proceso de TE salvo, como ya se ha comentado, las diferencias de procesamiento para la tarea de Recuperación de Información: no se trunca la lista de términos candidatos con criterios de precisión de la terminología extraída, sino que se mantienen todos los candidatos y el sistema recupera los más plausibles de acuerdo con la consulta.

5.3.1 Interfaz del primer prototipo

En la *Figura 5-7* puede observarse la forma en que se pide y presenta la información:

1. Un área de texto (área superior) permite al usuario introducir las palabras de búsqueda a partir de los cuales el sistema obtiene y presenta los dos rankings (documentos y términos).
2. En el área de documentos (área inferior izquierda) se muestra el código del documento con un enlace al mismo y se describe el documento por los términos que contiene, con enlaces a todos los contextos en los que interviene.

3. En el área de términos (área inferior derecha) se muestran los términos con un enlace a sus contextos, así como la lista de documentos (códigos numéricos) que contienen cada término y que se pueden acceder mediante un hiperenlace. La organización de los términos es una simple lista de sintagmas agrupados por sus componentes sin jerarquía alguna.

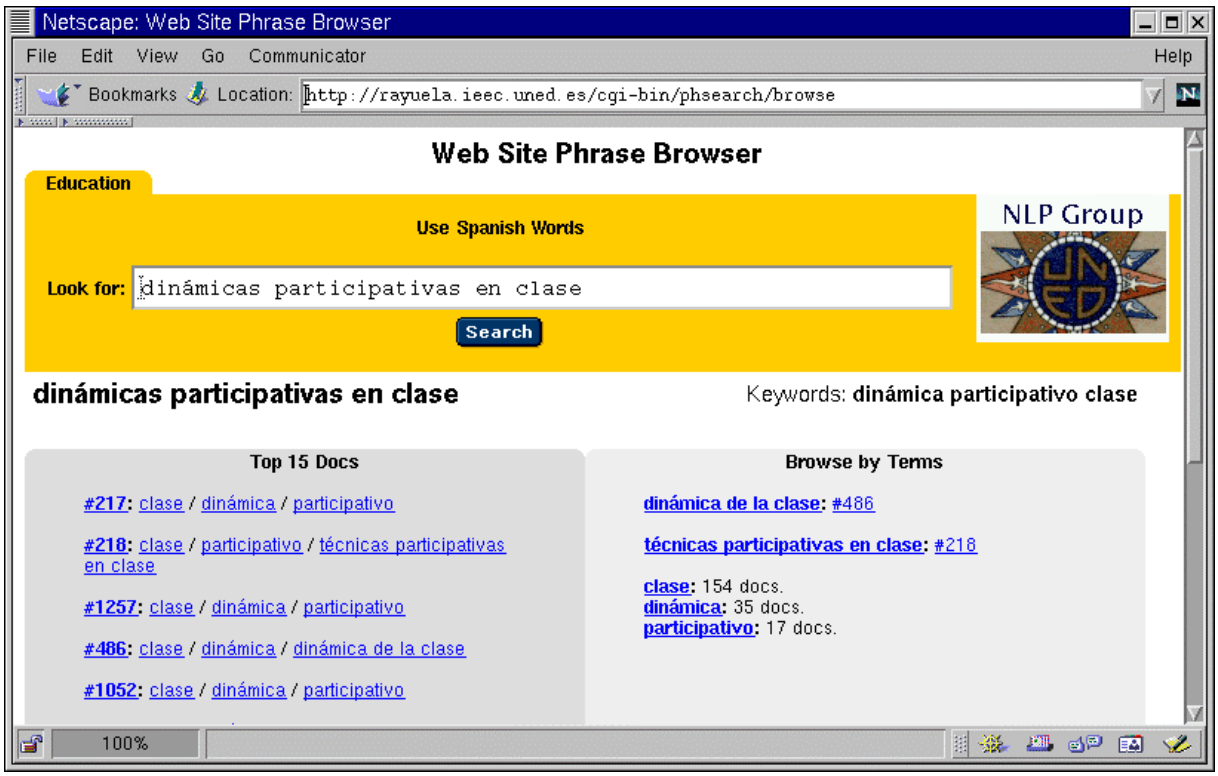


Figura 5-7. Interfaz del primer prototipo

Si el usuario pincha en un término (e.g. "participativo") se muestra la página de todos los contextos del término en la colección (*Figura 5-8*).

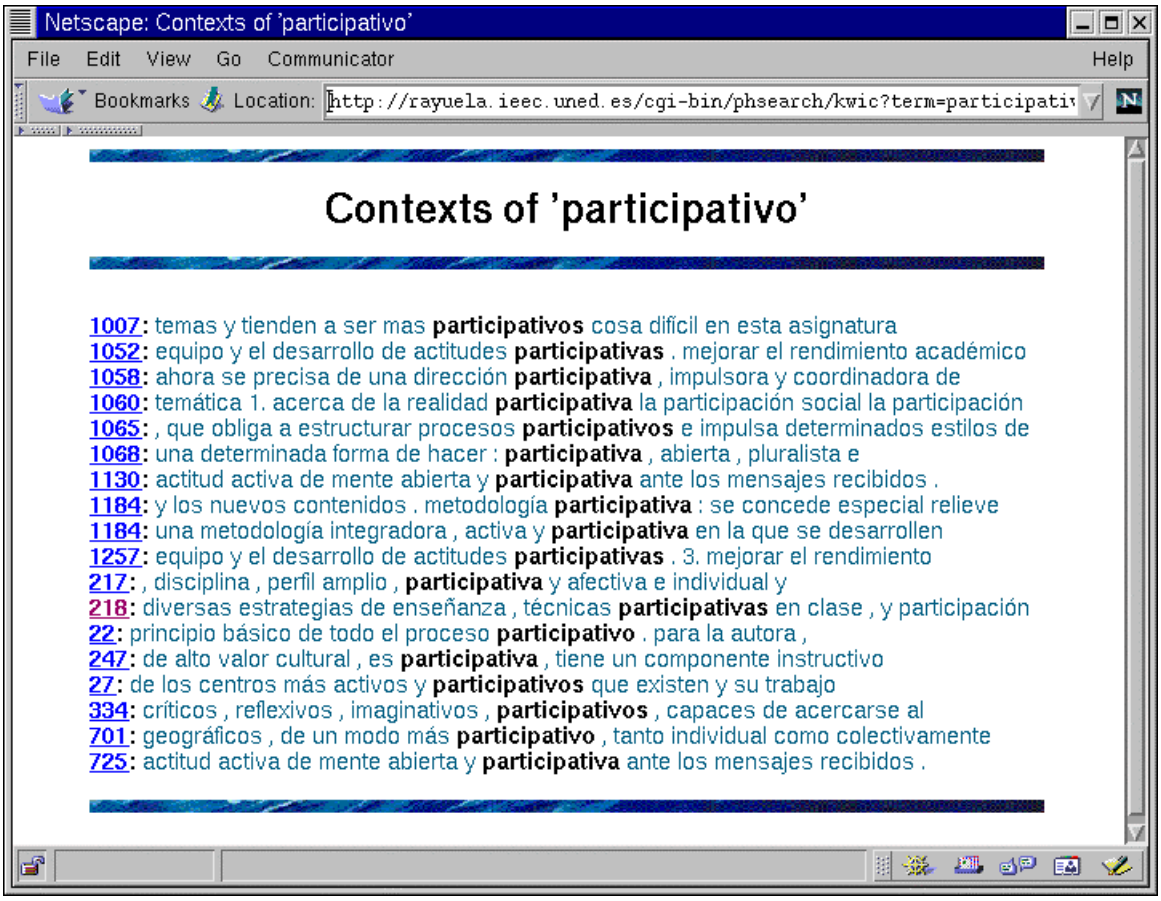


Figura 5-8. Contextos de un término en el primer prototipo

Cada uno de estos contextos tiene asociado un hiperenlace al documento que lo contiene. En el ejemplo, el usuario ha accedido al documento #218 (*Figura 5-9*).

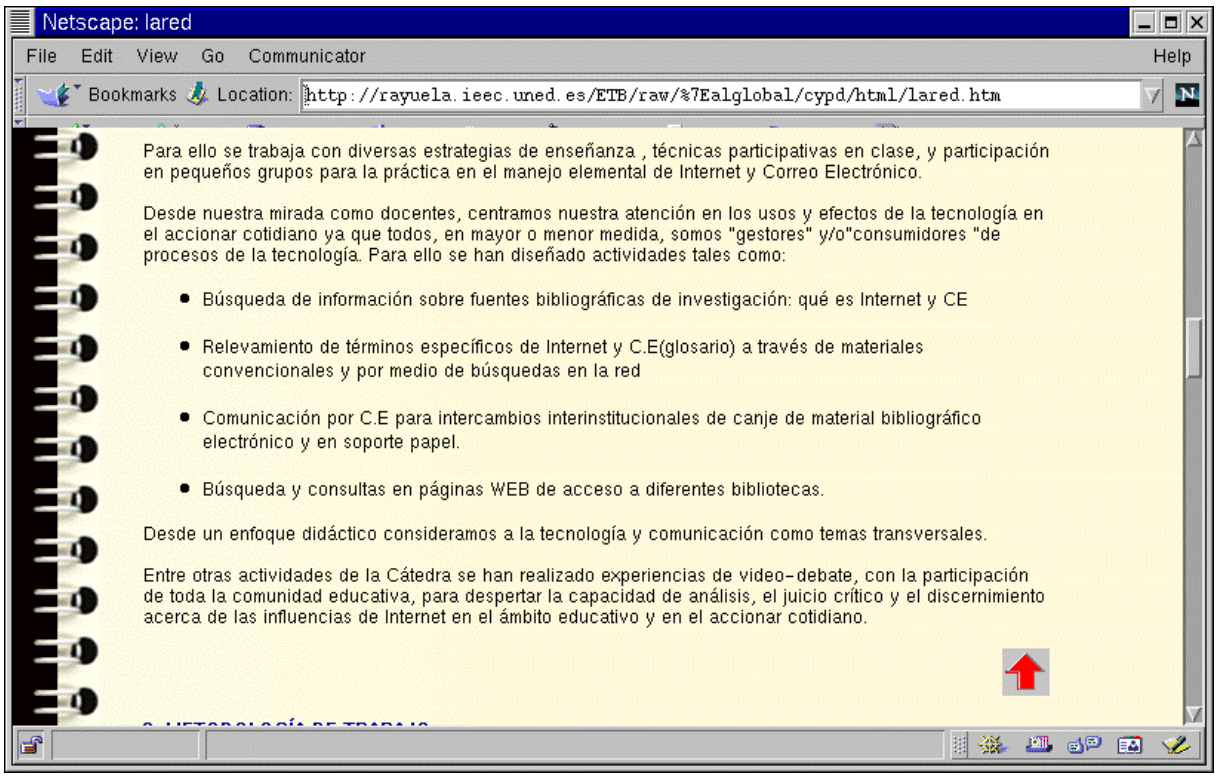


Figura 5-9. Documento número 218

5.3.2 Carencias detectadas en el primer prototipo

La restricción de co-ocurrencia de los lemas de la consulta en algún sintagma de la colección es muy fuerte, es decir, en muchas ocasiones la recuperación de sintagmas resulta infructuosa, y más en una colección tan pequeña. El siguiente prototipo abordará este problema expandiendo la consulta mediante lemas relacionados semánticamente con los lemas originales.

5.4 Segundo Prototipo

El objetivo del segundo prototipo es estudiar cómo afecta a la recuperación de sintagmas la expansión de la consulta mediante palabras sinónimas, hiperónimas e hipónimas identificadas mediante EuroWordNet a partir de la consulta. Asimismo, se plantea como objetivo estudiar de forma cualitativa el papel de la desambiguación de categoría gramatical y de las relaciones semánticas entre categorías(Peñas 2000;Chugur 2001).

Para este prototipo se ha utilizado la misma colección que en el primer prototipo, con el mismo proceso de extracción de sintagmas e indexación.

5.4.1 Desambiguación de la categoría gramatical

Respecto a la categoría gramatical de las palabras de la consulta, las alternativas que se presentan son desambiguar utilizando el etiquetador para el español (Carmona 1998) o no desambiguar, es decir, mantener todos los lemas de cada palabra de la consulta. En el segundo caso, la co-ocurrencia de lemas en un mismo sintagma tiene como efecto implícito la desambiguación del lema y categoría de las palabras de la consulta.

5.4.2 Expansión mediante EuroWordNet

Las opciones de expansión semántica vienen determinadas por las relaciones principales de EuroWordNet: sinonimia, hiperonimia e hiponimia. En cada uno de los tres casos, se da la posibilidad de expandir las palabras de la consulta mediante cualquiera de estas relaciones de forma individual o simultánea. El efecto de la expansión ya se ha descrito varias veces: proporcionar nuevas combinaciones de lemas que permitan recuperar un mayor número de sintagmas.

Las jerarquías de WordNet están asociadas a la categoría gramatical de las palabras, de forma que las jerarquías relativas a nombres, verbos, adjetivos y adverbios son independientes entre sí, no habiendo relaciones entre sus respectivos conceptos. Sin embargo, la relación entre palabras de distinta categoría puede ser muy estrecha. En el caso de los sintagmas dos expresiones pueden ser completamente equivalentes y diferenciarse únicamente en el uso de una preposición y un nombre en lugar del adjetivo correspondiente (e.g. “estudio demográfico” vs. “estudio de demografía”).

Esta carencia de WordNet se ha intentado paliar de forma parcial en EuroWordNet. Con el fin de considerar este tipo de variaciones morfosintácticas se aprovecharon los trabajos de (Gonzalo 1998a), (Peñas 2000) y (Chugur 2001) incorporando en el prototipo la posibilidad de expandir las palabras de la consulta mediante palabras de otras categorías (relaciones Cross-POS).

5.4.3 Interfaz del segundo prototipo

La novedad que incorpora el interfaz del segundo prototipo respecto al primero es la posibilidad de que el usuario decida:

1. Si se debe desambiguar o no la categoría de las palabras de la consulta. Si se opta por la desambiguación, la consulta se procesa mediante el etiquetador proporcionando un único lema para cada palabra original.
2. Qué expansiones semánticas quiere que se realicen: relaciones de sinonimia, hiperonimia, hiponimia y/o cross-pos.

Para ello, en el área de consulta del interfaz (área superior) se han habilitado una serie de *checkbox* que el usuario puede activar o desactivar. Las otras dos áreas (documentos en la inferior izquierda y términos la inferior derecha) permanecen igual. La *Figura 5-10* muestra el mismo ejemplo que en el interfaz anterior sólo que, en esta ocasión, se han activado las opciones de desambiguación de categoría y de expansión.

Figura 5-10. Interfaz del segundo prototipo

Puede observarse que la consideración de sinónimos da lugar a la recuperación de nuevos términos. En el ejemplo, *aula* es uno de los sinónimos de *clase* que ha dado lugar a la recuperación de un nuevo sintagma: *dinámica en el aula*. Este sintagma supone una variación tanto morfosintáctica como semántica de la consulta. Por otro lado, la consideración de los nuevos términos (*aula* y *dinámica en el aula*) ha dado lugar a un nuevo ranking de documentos.

5.4.4 Evaluación cualitativa

Con el fin de evaluar como afectaban a la recuperación de sintagmas las opciones descritas así como sus distintas combinaciones, se aplicaron al prototipo las mismas consultas pero con opciones diferentes.

Las conclusiones principales que se sacaron de la evaluación fueron:

1. Resulta preferible no desambiguar la categoría de las palabras de la consulta. Efectivamente, el coste de desambiguar la consulta mediante un etiquetador de categoría no se justifica, sino que al contrario, perjudica la recuperación de sintagmas por cerrar posibilidades que pueden ser correctas. Esto se debe principalmente a que la consulta es demasiado corta y esta falta de información hace que se produzcan errores en el etiquetado.
2. La expansión por hipónimos e hiperónimos aporta lemas que en ocasiones están demasiado alejados de la consulta original, y su combinación tiene como efecto recuperar muchos sintagmas sin relación relevante con la consulta. Es decir, salvo en casos concretos, difíciles de predeterminar, la expansión por hipónimos e hiperónimos introduce demasiado ruido.
3. Sin embargo, la expansión por sinónimos sí resulta de interés, permitiendo recuperar expresiones diferentes a la consulta, pero muy relacionadas con ésta e incluso equivalentes (e.g. “enseñanza a distancia” y “educación a distancia”, en el que *educación* es resultado de expansión por sinónimos del lema *enseñanza*). La proximidad de los sinónimos con respecto a los lemas originales de la consulta permite la recuperación de nuevos sintagmas sin apenas introducir ruido.
4. Respecto a las relaciones entre categorías, a pesar del número considerado (más de 5.000), han resultado demasiado pobres para el dominio de aplicación, aportando lemas de expansión en muy raras ocasiones.
5. Por último, se hizo evidente el mal uso (en general ignorándolas) de las opciones de expansión que hacían los usuarios no familiarizados con EWN. En el caso de conservar las opciones de expansión en prototipos posteriores,

habría que expresarlas en términos más cercanos al usuario (e.g. *términos más específicos*, *términos más generales*, etc.)

5.4.5 Carencias del segundo prototipo

A pesar de que 1.000 es un número reducido de documentos, el coste de etiquetar los textos resulta bastante elevado (varios días de procesamiento sobre un Pentium II a 300 MHz). Uno de los objetivos del trabajo es procesar colecciones suficientemente grandes y adaptar las técnicas lingüísticas para que esto sea posible. Por esta razón, los siguientes prototipos deben superar esta limitación.

Asimismo, el tratamiento de la consulta producto de su expansión también trae como consecuencia un incremento en el coste de procesamiento y un problema en los tiempos de respuesta.

Por último, la expansión de la consulta tiene como consecuencia la recuperación, en algunas ocasiones, de un número muy elevado de sintagmas cuya exploración puede resultar difícil. Resulta necesario mejorar la organización y jerarquización de los sintagmas.

5.5 Tercer prototipo

El objetivo del tercer prototipo es superar las carencias del prototipo anterior, aumentar la colección un orden de magnitud y abordar el problema del multilingüismo.

5.5.1 Mejora del coste computacional

Respecto al problema del excesivo coste computacional en el procesamiento tanto de la colección como de las consultas, una de las alternativas consideradas fue el paso a una arquitectura distribuida en la que los procesos residieran en máquinas diferentes y se comunicaran mediante CORBA. También se consideró la posibilidad de utilizar Prolog con predicados remotos. Las pruebas que se llevaron a cabo, así como la experiencia con los proyectos ITEM[11] y RILE[12], muestran que la

[11] http://sensei.lsi.uned.es/item
[12] http://rile.sema.es

arquitectura distribuida no es suficiente sin una mejora en el coste del procesamiento lingüístico.

Por esta razón, el problema de eficiencia se ha abordado:

1. Relajando el procesamiento lingüístico y adaptándolo a las necesidades concretas de recuperación de sintagmas y documentos, en la dirección que se ha explicado en el capítulo relativo al modelo propuesto (sección 4.2.2.3). En este prototipo aún no se realiza el etiquetado heurístico dirigido a la extracción de sintagmas (sección 4.2.2.4), sino que las palabras han sido etiquetados con su categoría más frecuente.
2. En el aspecto técnico, mediante la incorporación de bases de datos basadas en ficheros con modelo clave-valor (SGBD de Berkeley[13]). Tanto los índices como los recursos léxicos se adaptaron e introdujeron en estas bases de datos.

5.5.2 Expansión de la consulta

Respecto a la expansión de la consulta, la experiencia del prototipo anterior llevó a eliminar las opciones de expansión. Por un lado, los usuarios no aprecian este tipo de interacción y no disponen de información suficiente para saber a priori que elecciones realizar. Por otra parte, ni la desambiguación de categoría, ni la expansión por hiperónimos/hipónimos, ni por relaciones entre categorías permiten mejorar, en términos generales, la calidad de los sintagmas recuperados. Por estas razones, para el tercer prototipo se decidió no realizar desambiguación de categoría de las palabras de la consulta, y expandir únicamente por sinónimos, no dando opción al usuario de alterar este marco. De esta forma, vuelve a simplificarse el interfaz, se reduce la información que debe considerar el usuario, se deja al sistema las inferencias lingüísticas oportunas y únicamente se muestran los resultados de este procesamiento para que el usuario tome las decisiones finales en el acceso a la información.

5.5.3 Multilingüismo

Para este prototipo se utilizó la colección multilingüe de noticias internacionales, lo cual obligó a incorporar las herramientas para procesar inglés y catalán, así como los patrones para detectar los sintagmas en estos idiomas. Respecto al español, se siguieron utilizando los mismos patrones que en los prototipos anteriores, que son

[13] http://www.sleepycat.com

los utilizados en el trabajo preliminar de extracción de terminología. Estos mismos patrones fueron aplicados para el catalán mientras que en el caso del inglés se desarrollaron unos nuevos (*Figura 5-11*).

Español y catalán	Inglés
1. N N 2. N A 3. N [A] Prep N [A] 4. N [A] Prep Art N [A] 5. N [A] Prep V [N [A]]	1. A N [N] 2. N N [N] 3. A A N 4. N A N 5. N Prep N

Figura 5-11. Patrones morfosintácticos para la identificación de sintagmas terminológicos

5.5.4 Interfaz del tercer prototipo

El interfaz mantiene las tres áreas habituales: consulta, términos y documentos (*Figura 5-12*). El área de términos (inferior izquierda) incorpora unas carpetas desplegables que permiten organizar los sintagmas en una jerarquía de dos niveles. En el nivel superior de esta jerarquía (fondo oscuro) se presenta el representante del grupo y en el segundo nivel (fondo claro) todos los sintagmas miembros del grupo. Cada grupo corresponde a una combinación de los lemas de expansión y traducción de la consulta. La ordenación de los grupos y miembros de un grupo se realiza de la siguiente manera:

1. Los términos dentro de un grupo se ordenan por el número de documentos en los que aparece (como estimación del grado de lexicalización).
2. La ordenación de los grupos se realiza:
    a. En primer lugar, por el número de palabras relacionadas con la consulta que contienen los sintagmas (identificados como *Top Terms* si contienen más de dos).
    b. En segundo lugar, por el número de documentos diferentes que contienen a los miembros del grupo.

Esta ordenación no atiende al idioma de los términos por lo que aparecen los grupos de diferentes idiomas aparecen mezclados.

Figura 5-12. Interfaz del tercer prototipo (versión I)

Respecto al área de documentos (inferior derecha) se desarrollaron dos versiones. En la primera se muestran los documentos que contienen a un sintagma seleccionado sin ningún tipo de ordenación. Como información de ayuda a la discriminación de documentos se ofrece el contexto del sintagma en el documento (*Figura 5-12*).

En la segunda versión del área de documentos, éstos vienen ordenados y descritos por sus términos (palabras y sintagmas) relacionados con la consulta (*Figura 5-13*). El criterio de ordenación es el número de términos diferentes relacionados con la consulta y contenidos en el documento.

Figura 5-13. Interfaz del tercer prototipo (versión II)

5.5.5 Carencias del tercer prototipo

Respecto a la exploración de sintagmas, los criterios de ordenación de términos funcionan bien cuando hay relevancia estadística, es decir, cuando el número de ocurrencias de los sintagmas es suficiente (al menos 2 ó 3 ocurrencias en documentos diferentes). Sin embargo, en grupos dónde los sintagmas únicamente aparecen en un solo documento la elección del representante del grupo resulta arbitraria. Por otra parte, es necesario recoger la relación de inclusión de sintagmas. Subsintagmas de un mismo sintagma se vuelven a presentar aunque aparezcan en los mismos documentos.

Otro problema se debe a que la cobertura de EWN es pobre, sobre todo en lo referente a adjetivos. Es necesario utilizar recursos adicionales como diccionarios bilingües y tratar de establecer relaciones entre categorías. El enriquecimiento de

EWN en este sentido, no es suficiente tal como se apreció en el segundo prototipo, por lo que una aproximación posible sería buscar algún tipo de convergencia entre los lemas, más que a considerar derivaciones nominales o adjetivales.

Por último, en este prototipo se hizo evidente que un alto porcentaje de los usuarios no ponen tildes en las palabras de la consulta. Esto, en el caso del español, resulta una práctica bastante habitual y un problema debido a que no se puede acceder a las entradas correspondientes de los diccionarios. La eliminación de caracteres especiales (incluidas las tildes) requiere una adaptación de todos los recursos lingüísticos.

5.6 Cuarto prototipo

El objetivo del cuarto prototipo es abordar las limitaciones del prototipo anterior, así como estudiar la escalabilidad del sistema a nuevos idiomas y a colecciones más grandes. Para ello se ha utilizado la colección multilingüe de recursos educativos en cinco idiomas: español, inglés, francés, italiano y catalán.

La cercanía de estas lenguas permite generalizar los patrones de extracción de sintagmas. De esta forma, como ya se ha descrito y justificado en el modelo propuesto (4.2.2.2), la mayoría de sintagmas nominales terminológicos responden a un mismo patrón general. En este prototipo se ha incorporado el etiquetado heurístico a partir del análisis morfológico de las palabras, dirigido a la extracción de sintagmas para el español y catalán.

5.6.1 Incorporación de nuevos idiomas

La adición de nuevos idiomas al sistema depende de los recursos y herramientas disponibles para cada uno de ellos. Así, por ejemplo, disponiendo de un etiquetador apropiado, la adición del francés y el italiano a los tres idiomas ya contemplados (español, inglés y catalán) resulta relativamente sencilla pues son lenguas cercanas y responden a las mismas reglas generales.

No ocurre lo mismo con el alemán que, como muestran las experiencias en CLEF 2000, requieren herramientas específicas de segmentación de palabras en sus componentes. Por esta razón y aunque el alemán está presente en la colección construida para este prototipo, los documentos en este idioma han tenido que ser ignorados.

Respecto al coste de procesamiento de la consulta, éste crece linealmente con el número de idiomas, pues el proceso de traducción debe repetirse para cada uno de ellos. Sin embargo, el coste de extracción de términos e indexación de la colección permanece igual, dependiendo únicamente del número de documentos, no de los idiomas contemplados. Esto es así porque se identifica el idioma de los documentos y se realiza un procesamiento separado para cada idioma.

5.6.2 Adaptación e incorporación de recursos

Al margen de los recursos y herramientas que debieron incorporarse para los nuevos idiomas, la falta de cobertura de EWN llevó a la incorporación de diccionarios bilingües, aunque sólo disponibles para el par inglés-español-inglés.

Todos los recursos (EWN, diccionarios y tablas de lemas) fueron adaptados para que sus entradas no dependieran del uso correcto de caracteres especiales, en especial eliminando las tildes de las palabras. Esto requiere también eliminar las tildes en las palabras de la consulta y asegurar la consistencia en las indexaciones de la colección, tanto de documentos como de sintagmas. De esta forma se uniformizó el comportamiento del sistema haciéndolo capaz de procesar cualquier consulta independientemente del uso de caracteres especiales que hicieran los usuarios.

La eliminación de caracteres especiales introduce una componente de ambigüedad en el sistema, pues se reducen palabras diferentes a formas comunes. Sin embargo, la restricción que los sintagmas imponen sobre las palabras que los componen vuelve a deshacer la ambigüedad introducida.

5.6.3 Interfaz del cuarto prototipo

El interfaz del cuarto prototipo es muy similar al del prototipo anterior. La única novedad viene determinada por la incorporación de nuevos idiomas. Así, a diferencia de los prototipos anteriores en los que se mezclaban sintagmas de diferentes idiomas, el nuevo interfaz explora la posibilidad de separar los sintagmas por idiomas, ofreciéndolos al usuario en carpetas diferentes (*Figura 5-14*).

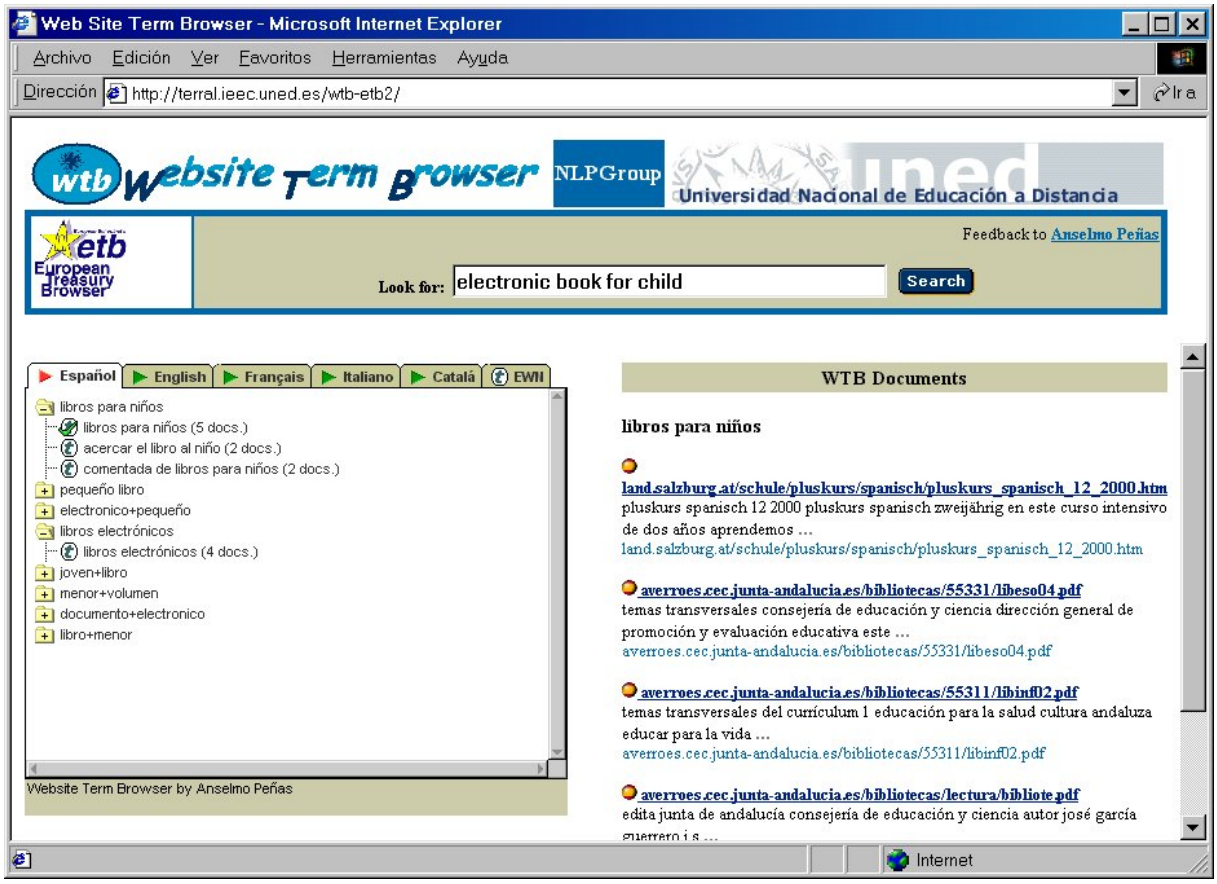


Figura 5-14. Interfaz del cuarto prototipo

5.6.4 Carencias del cuarto prototipo

Una vez superados los problemas de eficiencia y abordados en la medida de lo posible los problemas de cobertura el sistema se comporta razonablemente bien con colecciones equivalentes a sitios web grandes (decenas de miles de documentos), tanto en lo que respecta a los tiempos de respuesta como a los términos que propone al usuario en los cinco idiomas.

Sin embargo, la evaluación del sistema debe realizarse no sólo en términos cualitativos sino también cuantitativos. Así pues, es necesario diseñar un marco de evaluación que determine el desarrollo de un último prototipo sobre el que realizar la evaluación cuantitativa del modelo propuesto. En este sentido, quedan pendientes de evaluación varios aspectos:

1. La utilidad de la exploración de sintagmas en el acceso a la información.

2. La capacidad del modelo para realizar un acceso translingüe a la información.
3. La capacidad del modelo para indexar colecciones grandes.

El primer aspecto, la utilidad de los sintagmas, conduce al desarrollo del siguiente prototipo. La evaluación del segundo aspecto, capacidad de acceso translingüe a la información, se ha realizado sobre este cuarto prototipo y se muestra en el próximo capítulo de evaluación (6.3). El tercer aspecto, la capacidad de tratar colecciones grandes, no ha podido evaluarse con este cuarto prototipo. La falta de herramientas lingüísticas para el alemán ha llevado a descartar un porcentaje elevado de documentos dejando la colección en poco más de la mitad de los 100.000 especificados para este prototipo. Esto ha motivado un trabajo adicional que se muestra también en el siguiente capítulo de evaluación (**¡Error! No se encuentra el origen de la referencia.**).

5.7 Quinto prototipo

De acuerdo con las especificaciones, el objetivo del quinto prototipo es establecer una forma de evaluación cuantitativa de la utilidad de explorar sintagmas en el acceso a la información. Además, se estudiarán vías alternativas de organización y utilización de los términos en el acceso a la información.

Este quinto prototipo se ha aplicado a dos colecciones, (i) la misma colección que el prototipo anterior (colección multilingüe de recursos educativos) y (ii) a la colección de páginas web en el dominio UNED.es. Esta última colección se dirige a una comunidad muy concreta (personal y alumnos de la UNED) con unas necesidades reales de información. De esta forma, la implantación del sistema en un entorno con necesidades reales de información garantiza que en unos meses se hayan realizado varios miles de consultas sobre las que realizar la evaluación del sistema.

5.7.1 Organización de los sintagmas

La organización de los sintagmas propuesta en este prototipo ya se ha discutido en el modelo propuesto. Una recuperación translingüe, en general, se justifica en el caso de que la información que se busca no se encuentre ya en el mismo idioma de la consulta. Se establece, así, un criterio de ordenación de la terminología recuperada por proximidad a la forma y el idioma de la consulta. En primer lugar se ofrecen al usuario los sintagmas con los mismos lemas y el mismo idioma que la consulta original. Posteriormente se sugieren al usuario los sintagmas que

contienen sinónimos o traducciones de los lemas originales de la consulta. Los sintagmas en otros idiomas, por tanto, aparecerán en último lugar puesto que son los más lejanos a la forma de la consulta. El supuesto que se maneja es que el usuario es el responsable de elegir las palabras que él considera más adecuadas para identificar la información que busca. Un sistema de búsqueda no puede ignorar este principio, sino que debe considerar las elecciones del usuario al realizar la consulta.

5.7.2 Recuperación de documentos mediante Google

La mayoría de los sistemas de búsqueda ofrecen como resultado una lista de los documentos más relevantes de acuerdo con la consulta. Google (http://www.google.com) es uno de los mejores buscadores de este tipo accesibles en Internet. Su ordenación de documentos basada, entre otros criterios, en la topología de Internet, proporciona excelentes resultados. La pregunta que se plantea es la siguiente: si los sistemas como Google funcionan tan bien, ¿resulta útil o no la incorporación del área adicional de términos propuesta en este trabajo?. Es decir, ¿aporta algo la recuperación y propuesta de términos a la recuperación y ordenación de documentos? ¿Aprecian los usuarios una información como la que presenta Website Term Browser?

De esta forma, se plantea la comparación entre el uso del área de términos y el área de documentos pero, para que la información presente en el área de documentos esté fuera de duda en cuanto a su calidad, la propuesta de términos de WTB se compara con la propuesta de documentos de Google sobre el mismo dominio.

La comparación entre los dos sistemas sobre el mismo dominio de la UNED es posible gracias a que Google tiene perfectamente indexado este dominio y a que Google permite restringir una consulta a un determinado dominio (en este caso UNED.es). Es necesario observar que la indexación de WTB y la de Google son diferentes, pero que estas diferencias, en todo caso, benefician a Google pues su indexación y actualización de los documentos indexados es mejor y más exhaustiva que la de WTB. Esto es así porque:

1. Actualiza con frecuencia la indexación incorporando las modificaciones de las páginas en el dominio.
2. Al considerar toda la web, Google tiene más puntos de acceso al dominio. Algunos servidores de la UNED no están referenciados entre sí. Las páginas del dominio no tienen estructura de árbol sino de grafo inconexo porque existen servidores propios de los departamentos y grupos de investigación no referenciados por las páginas del servidor principal. La posibilidad de acceder

a estas páginas desde sitios externos al dominio uned.es hace más rica la indexación del dominio.

Así pues, las diferencias de indexación de los dos sistemas no suponen un elemento de distorsión para la evaluación de WTB, en el sentido de que pudieran falsear los resultados en beneficio de una evaluación positiva de WTB.

5.7.3 Re-consulta con un sintagma

Puesto que el área de documentos pasa a ser alimentada tanto por WTB como por Google, dependiendo de las acciones del usuario, cabe una nueva forma de utilizar los sintagmas recuperados por WTB: utilizarse como una nueva consulta a Google. Así, el usuario puede seleccionar uno de los sintagmas propuestos por WTB con dos finalidades, o bien ver los documentos indexados por WTB que contienen dicho sintagma, o bien lanzar una nueva consulta a Google utilizando el sintagma como contenido de la consulta. Si con la primera acción se obtiene una lista de documentos proporcionada por WTB, la segunda tiene como efecto la modificación del área de documentos con el ranking de documentos que da Google como resultado a la nueva consulta. De esta forma, se utilizan los sintagmas a modo de *relevance feedback* para reconsultar a Google.

5.7.4 Registro de la interacción

Para llevar a cabo la evaluación del sistema sobre la base de su utilización en un entorno real de trabajo, ha sido necesario añadir a WTB un módulo cuya función es el registro de las interacciones de los usuarios. La interacción de los usuarios se registra en forma de una *sesión* que comienza con una consulta y prosigue con la secuencia arbitraria de las siguientes acciones:

1. Listar los documentos que contienen a un determinado término (*explorar término*).
2. Utilizar un sintagma como nueva consulta a Google (*re-consultar*).
3. Explorar un documento.

En el registro de una sesión, junto con el tipo de acción que ha realizado el usuario se almacenan en un fichero datos como la consulta realizada, términos seleccionados, documentos explorados, etc.

5.7.5 Interfaz del quinto prototipo

La *Figura 5-15, Figura 5-16, Figura 5-17 y* Figura 5-18 muestran el interfaz del quinto prototipo sobre la *colección de páginas web del dominio UNED*.

Tras la consulta del usuario, WTB ofrece al usuario las dos áreas habituales, la de términos y la de documentos. El primer ranking de documentos que se ofrece es el que proporciona Google como resultado de transmitirle directamente la consulta (*Figura 5-16*). El usuario puede seleccionar un documento de esta lista para explorarlo, o bien puede seleccionar un término del área de términos con dos finalidades diferentes:

1. Obtener la lista de documentos indexados por WTB que contienen al término seleccionado (*Figura 5-17*).
2. Reconsultar a Google con un sintagma, con el resultado de obtener un nuevo ranking de Google en el área de documentos (*Figura 5-18*).

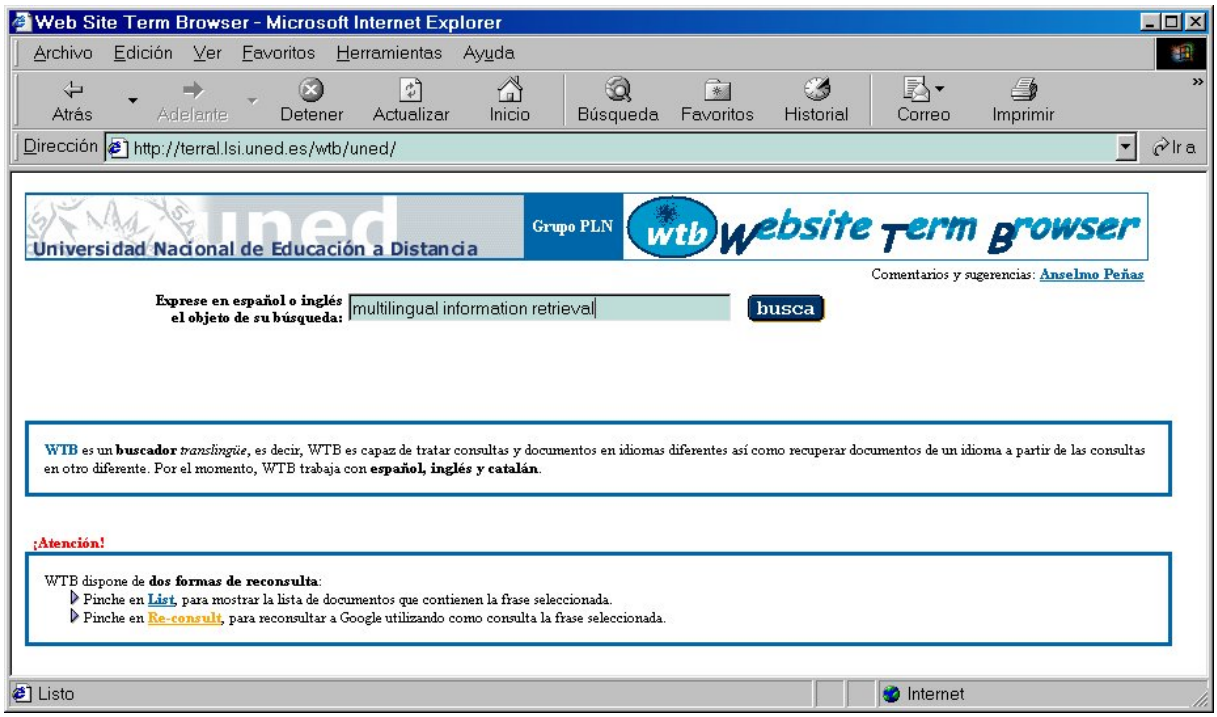


Figura 5-15. Interfaz del quinto prototipo, página de entrada.

En la *Figura 5-16* puede observarse que *"multilingual information retrieval"* no es un término que se utilice en la colección (si fuera así, aparecería en primer lugar en el área de términos). El término que sí aparece en la colección es *"multilingual text retrieval"*, que aparece en primer lugar. Este término se utiliza en la *Figura 5-17*

para listar los documentos que lo contienen, y se utiliza en la *Figura 5-18* como nueva consulta a Google.

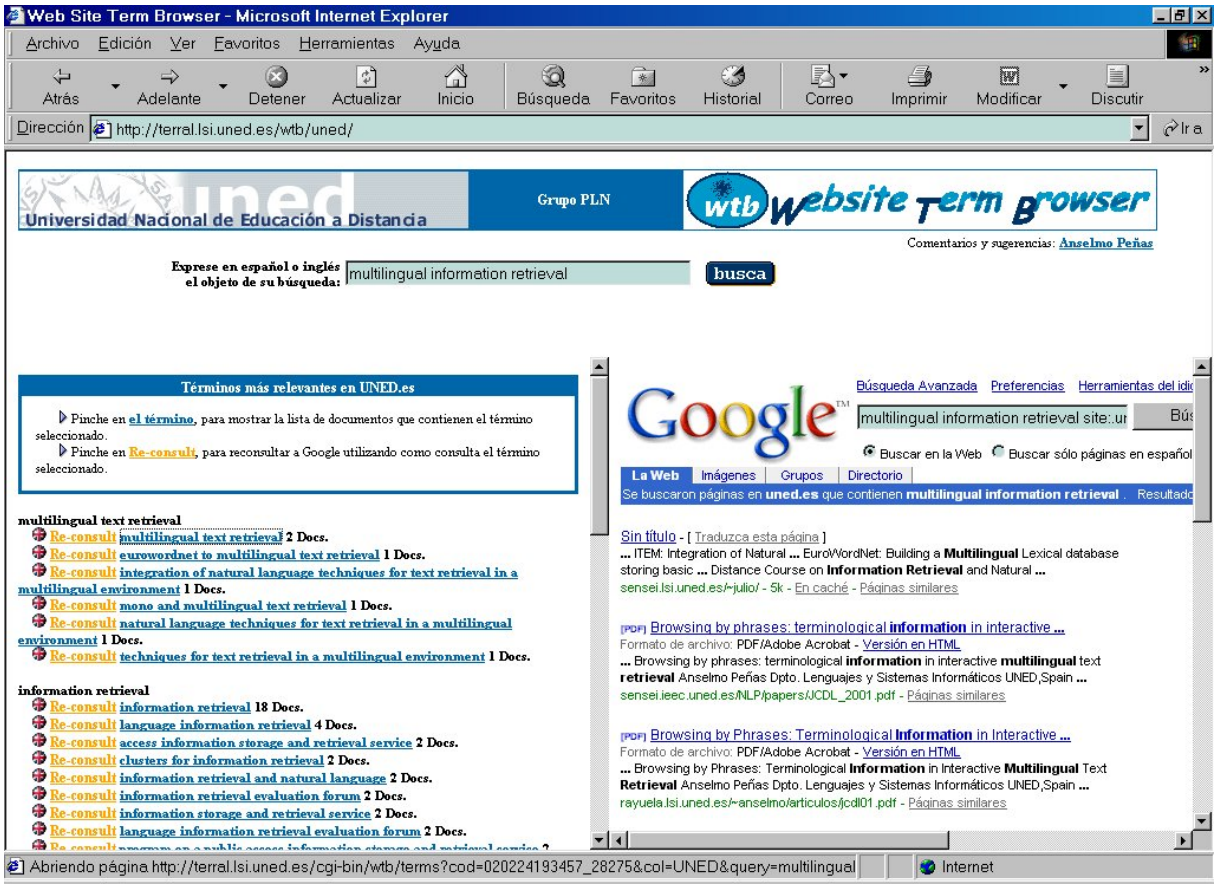


Figura 5-16. Resultado inicial de una consulta.

En el área de términos pueden observarse otros términos que pueden ayudar a precisar la información que busca el usuario:

- eurowordnet to multilingual text retrieval
- integration of natural language techniques for text retrieval in a multilingual environment
- clusters for information retrieval
- information retrieval and natural language
- information retrieval evaluation forum
- etc.

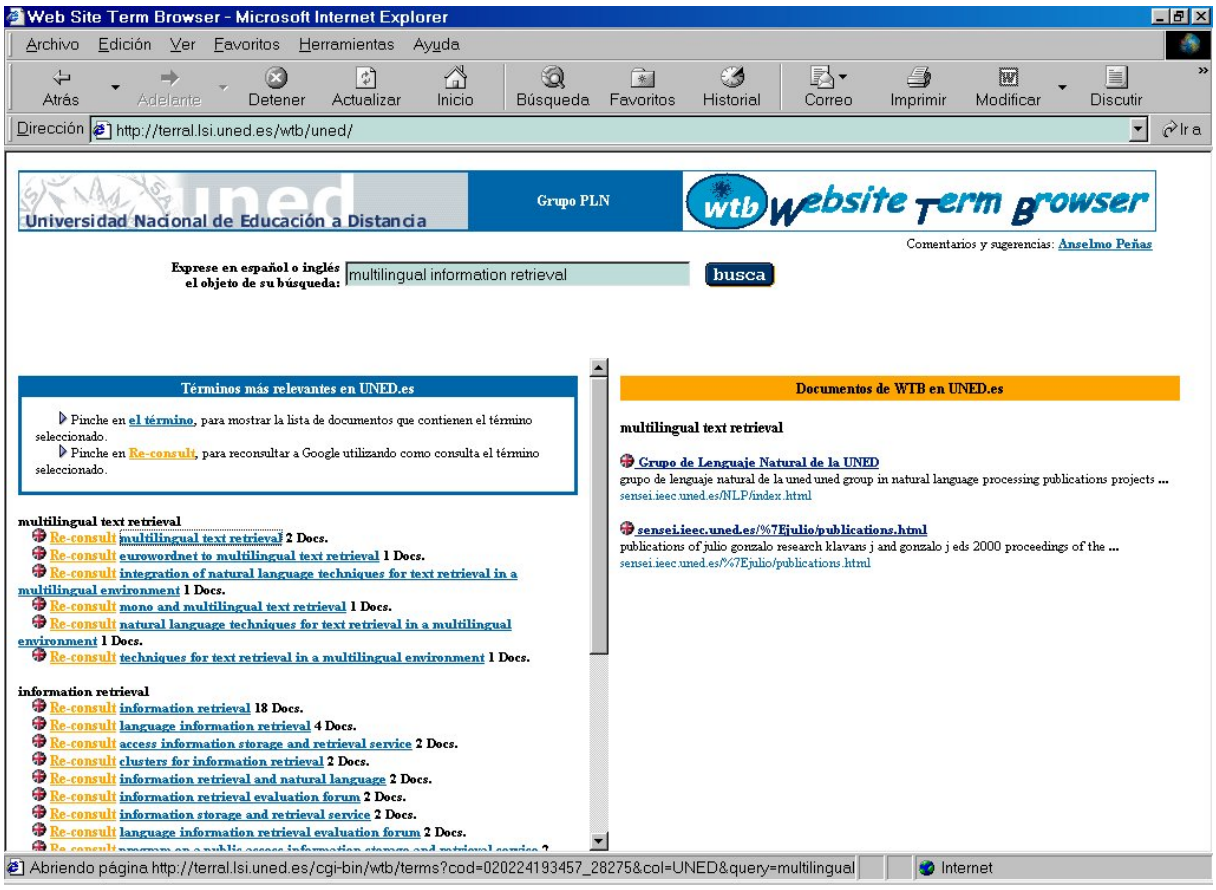


Figura 5-17. Explorar un término.

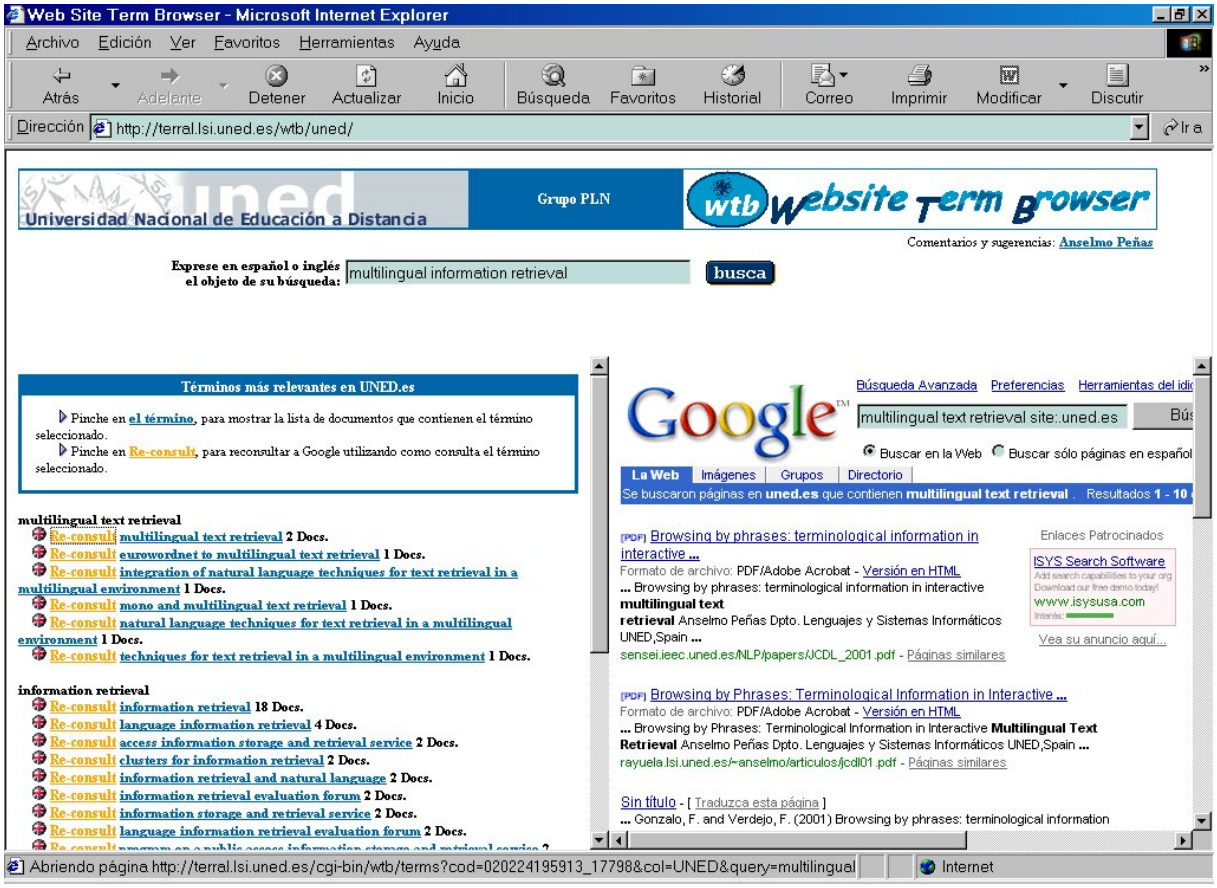


Figura 5-18. Re-consultar con un término.

La *Figura 5-19* muestra el interfaz del quinto prototipo sobre la *colección multilingüe de recursos educativos.* El usuario ha escrito una consulta en español *("escuela primaria y secundaria")* y el sistema le ha devuelto la lista de términos relacionados con la consulta. Primero se intentan encontrar aquellos términos más cercanos a las palabras que ha utilizado en la consulta, es decir, se premia la recuperación monolingüe. Sin embargo, en este caso la información solicitada no se encontraba en español y el sistema ha recuperado en primer lugar el término inglés "primary and secondary schools". El usuario ha seleccionado este término y el efecto ha sido mostrar la lista de documentos indexada por WTB que lo contienen.

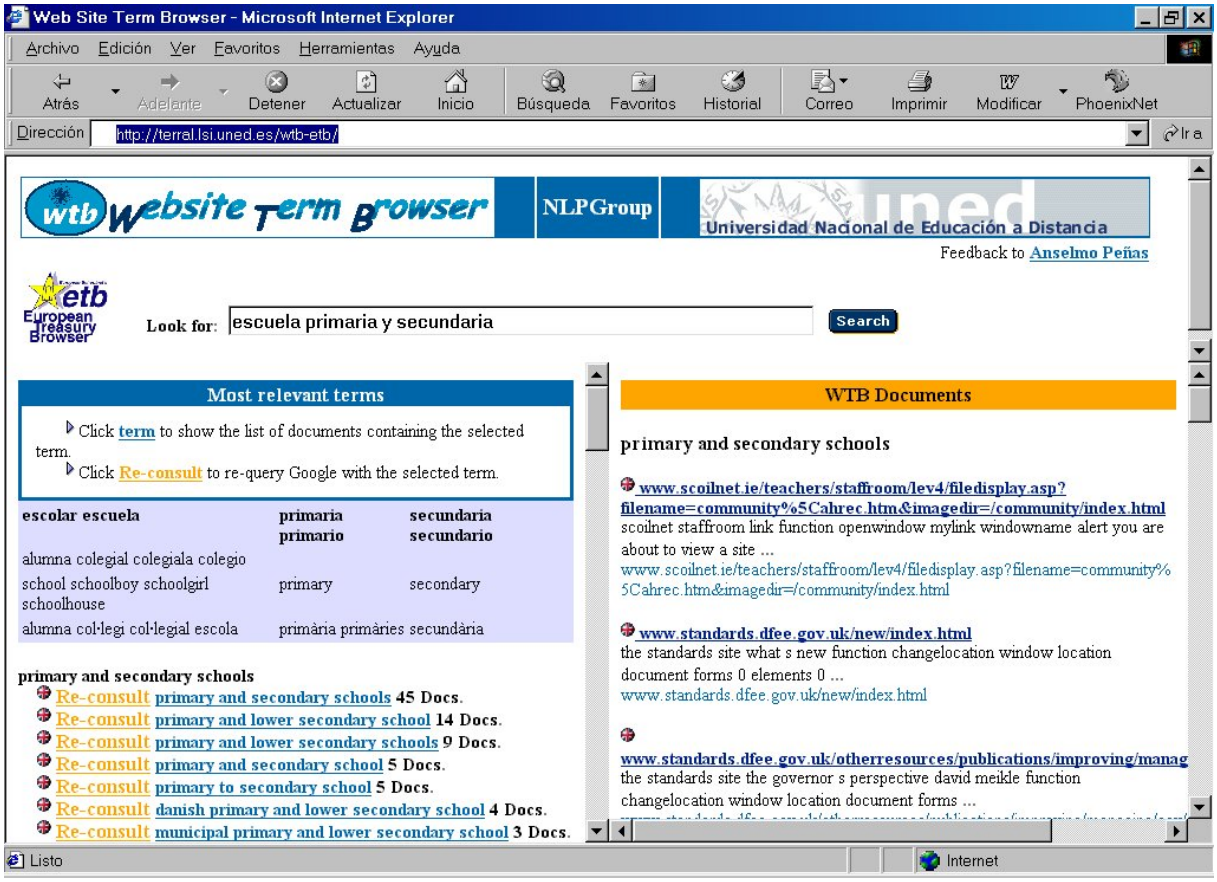


Figura 5-19. Interfaz del quinto prototipo sobre la colección multilingüe de recursos educativos

Capítulo 6

Evaluación

En este capítulo se recogen todos los aspectos de evaluación del modelo propuesto. El primer apartado del capítulo (6.1) discute las dificultades de evaluar un sistema interactivo de acceso a la información. Los marcos de evaluación existentes no son satisfactorios ni aplicables al modelo propuesto en este trabajo, por lo que el resto del capítulo presenta un marco novedoso de evaluación dirigido a obtener evidencias sobre la utilidad y capacidades del sistema desarollado.

La comparación entre el uso que hacen los usuarios del nuevo área de términos respecto al uso que hacen del ranking tradicional de documentos permite evaluar la utilidad de explorar sintagmas en recuperación de información (6.2). Esta evaluación se ha realizado en un entorno real de trabajo (la UNED) en el que los usuarios tienen necesidades reales de información y desconocen que su interacción con el sistema sirve para evaluarlo.

El tercer apartado (6.3) evalúa la capacidad del sistema para recuperar terminología en otros idiomas de forma translingüe. Para ello, se han utilizado como consultas a WTB los términos en cuatro idiomas de un tesauro multilingüe. Los términos recuperados por WTB en los cuatro idiomas se han comparado con los términos preferidos en el tesauro para obtener cotas inferiores de precisión y cobertura de la recuperación translingüe de terminología. Esta evaluación permite evaluar, además, la calidad de los recursos y herramientas de procesamiento lingüístico de los que depende el sitema.

El capítulo termina con la descripción de otras tareas de aplicación de WTB. La sección (6.4) describe la participación en el apartado interactivo de CLEF 2001 de un sistema de selección translingüe de documentos basado en los sintagmas extraídos por WTB. Por último, la sección (6.5) describe brevemente otras tareas en las que WTB puede resultar de utilidad.

6.1 Dificultades en la evaluación de la interactividad

La forma de evaluar un sistema tradicional de recuperación de documentos no puede aplicarse a Website Term Browser ya que requiere que el sistema devuelva una lista de documentos relacionados con la consulta y ordenada por criterios de relevancia. Sobre esta lista ordenada, y de acuerdo a los juicios de relevancia previamente asignados a la colección de prueba, se obtienen las medidas de precisión y cobertura de la recuperación para una serie de consultas. Esta forma de evaluar se ajusta a un modelo de recuperación de documentos en la que los usuarios no intervienen en el proceso de acceso y selección de información.

Las evaluaciones diseñadas con anterioridad para los sistemas de recuperación con algún grado de interactividad tampoco se ajustan a las características de WTB. El apartado interactivo de las conferencias TREC es una de las referencias más importantes a este respecto. Sin embargo, la necesidad de establecer criterios objetivos de comparación de los sistemas obliga a trabajar sobre consultas bastante precisas, en condiciones de laboratorio y con usuarios controlados. Sin embargo, los sistemas interactivos resultan de mayor utilidad cuando las consultas no son totalmente cerradas. El efecto de estos condicionantes es que no se encuentren diferencias apreciables entre los sistemas evaluados y que, por tanto, no se puedan determinar que elementos de interacción funcionan mejor.

Tampoco las evaluaciones diseñadas en la actualidad para sistemas interactivos de recuperación translingüe de información resultan satisfactorias para evaluar a WTB. Estas evaluaciones se dirigen a determinar:

1. Cómo los sistemas ayudan al usuario en la elección de los términos adecuados de consulta cuando el usuario no está familiarizado con el idioma de la colección (Ogden 1999) (Oard 2001).
2. Cómo los sistemas muestran los documentos escritos en lengua extranjera de forma que el usuario pueda juzgar su relevancia sin conocer el idioma (Oard 2001).

Ninguno de estos dos casos se ajusta a las características de WTB por lo se ha hecho necesario diseñar un nuevo marco de evaluación.

6.2 Evaluación de la utilidad del área de términos

En esta evaluación se ha utilizado el quinto prototipo. Su objetivo principal es evaluar la utilidad del sistema y, en concreto, la utilidad de presentar al usuario el área de términos.

6.2.1 Evaluación por comparación

Tal como se ha descrito anteriormente, el interfaz de WTB presenta dos tipos de información al usuario: un área de términos (palabras y sintagmas) y un área de documentos relevantes para la consulta. La hipótesis utilizada en la evaluación es que si el ranking de documentos es suficientemente bueno, el usuario seleccionará directamente uno de los documentos que se le ofrecen, y que decidirá utilizar la información terminológica cuando el ranking de documentos no satisfaga directamente sus necesidades de información y alguno de los términos propuestos por el sistema le parezca interesante. De esta forma, se habilita una forma de evaluación basada en la comparación del uso de ambos tipos de información, términos y documentos.

6.2.2 Evaluación en entorno real de trabajo

La evaluación planteada requiere el registro de las interacciones del usuario con el sistema, pero tiene como virtud que no se le pide al usuario un feedback sobre algo que no resulta de su interés: la evaluación del sistema que utiliza como herramienta. Esto permite que la evaluación pueda realizarse en condiciones reales de trabajo, lo cual supone una novedad significativa y una de las aportaciones importantes del trabajo.

El entorno real de trabajo que se ha seleccionado para la evaluación del sistema es el dominio UNED.es de la Universidad Nacional de Educación a Distancia. Este dominio posee más de 40.000 documentos accesibles a través de Internet y tiene asociado una población numerosa de usuarios (profesores, alumnos, personal administrativo, etc.) interesados en el acceso eficaz a toda esa información.

6.2.3 Comparación con los sistemas de búsqueda de documentos

La aportación de WTB es la sugerencia al usuario de términos que incluyen variaciones morfosintácticas, semánticas y translingües de la consulta, como información complementaria al ranking tradicional de documentos. Puesto que la evaluación comparativa se realiza respecto a este ranking, es necesario asegurar un buen ranking que proporcione objetividad en la evaluación de la utilidad del área de términos respecto al de documentos. Este ranking ha sido el proporcionado por Google[14] para el dominio UNED.es. Google es uno de los mejores buscadores existentes en Internet, su recolección de documentos es muy exhaustiva (mejor que la de WTB) y sus criterios de ranking basados en la topología de la red son los que mejores resultados están dando en términos de satisfacción del usuario.

6.2.4 Juego de acciones disponibles para el usuario

Tras la consulta inicial, la interfaz muestra, por una parte, un área de términos proporcionados por WTB y, por otra parte, un ranking de documentos proporcionados por Google. De esta forma, el usuario puede realizar las siguientes acciones arbitrariamente:

1. CONSULTA: escribir una nueva consulta.
2. EXPLORAR DOCUMENTO: seleccionar un documento para ver su contenido.
3. EXPLORAR TÉRMINO: seleccionar un término (fundamentalmente sintagmas) para ver la lista de documentos relacionados con él.
4. RECONSULTAR CON TÉRMINO: consultar de nuevo a Google utilizando como consulta un término (generalmente sintagma) seleccionado entre los que ha mostrado WTB.

Si el usuario selecciona un documento tras la consulta o una reconsulta, el documento habrá sido propuesto por Google. Si el usuario selecciona un documento tras la exploración de un término, el documento habrá sido propuesto por Website Term Browser.

6.2.5 Registro de la interacción de los usuarios

Todas las interacciones de los usuarios se registran en forma de *sesiones*. Una sesión comienza con la acción de *CONSULTA* y continúan con cualquier

[14] Google: http://www.google.com

combinación de acciones de *EXPLORAR DOCUMENTO, EXPLORAR TÉRMINO y RECONSULTAR CON TÉRMINO*. Las sesiones terminan cuando el usuario abandona el sistema o escribe una nueva consulta.

La *Figura 6-1, Figura 6-2 y Figura 6-3* muestran tres ejemplos de sesiones registradas en el sistema. La primera de ellas es una sesión con interacción en la que el usuario ha introducido una consulta *"ozone hole"*, después ha explorado un término *"degradación de la capa de ozono"* listando los documentos que lo contienen, y por último, ha explorado un documento de la lista.

```
LOG FILE 539

2001/03/14 12:10:33  CONSULTA    UNED   193.146.241.164   ozone hole
2001/03/14 12:11:20  EXPLORAR_TERMINO   539684: degradación de la capa de ozono
2001/03/14 12:11:29  EXPLORAR_DOCUMENTO  http://www.uned.es/doctorado/0108.htm
```

Figura 6-1. Sesión con interacción

La *Figura 6-2* muestra el fichero correspondiente a una sesión sin interacción, es decir, una sesión en la que no se ha realizado acción alguna aparte de introducir la consulta.

```
LOG FILE 010316120316_21841

2001/03/16 12:03:16  CONSULTA    UNED    62.204.196.27   american soldier
```

Figura 6-2. Sesión sin interacción

Por último, la *Figura 6-3* muestra una petición de búsqueda sin haber introducido la consulta oportuna, es decir, una sesión vacía.

```
LOG FILE 760

2001/03/14 17:11:44   EMPTY    UNED   212.128.26.72
```

Figura 6-3. Sesión vacía, sin consulta

La *Tabla 6-1* muestra el resumen de las estadísticas de interacción. Se han registrado en el sistema un total de 4731 sesiones, de las cuales el 4,7% están vacías, el 46,3% no tienen interacción a excepción de la consulta y el 49% restante corresponde a sesiones con interacción (2318 sesiones). Una de las razones de que hayan tantas sesiones sin interacción es que al pinchar en el botón de búsqueda algunos usuarios hacen doble click, con el efecto de abrir dos sesiones para la consulta en vez de sólo una.

La tabla muestra además que sólo en el 74,6% de las sesiones con interacción se llega a explorar un documento. En este sentido, en el 64,71% de las sesiones el usuario ha explorado al menos un término y en el 16% lo ha utilizado como nueva consulta a Google.

El último dato que muestra la tabla es el número medio de acciones por sesión. Además de la consulta, una sesión contiene en media algo menos de dos exploraciones de algún término, prácticamente dos exploraciones de algún documento y menos de una acción de reconsulta utilizando un término. En total, contando la consulta, una sesión contiene una media de 5 interacciones.

Sesiones sin consulta (vacías):	223/4731	5%
Sesiones sin interacción:	2190/4731	46%
Sesiones con interacción:	2318/4731	49%

EXPLORAR DOCUMENTO se usa en 1730/2318 sesiones:	75%
RECONSULTAR CON TÉRMINO se usa en 371/2318 sesiones :	16%
EXPLORAR TÉRMINO se usa en 1500/2318 sesiones:	65%

Número medio de acciones por sesión:

EXPLORAR DOCUMENTO 4539/2318:	1’958
RECONSULTAR CON TÉRMINO 804/2318:	0’346
EXPLORAR TÉRMINO 4301/2318:	1’855

Tabla 6-1. Resumen de datos de interacción

6.2.6 Secuencias de interacción más frecuentes

Las secuencias de acciones más frecuentes fueron:

```
2190 CONSULTA
 516 CONSULTA DOC
 288 CONSULTA TERM
 223 VACIA
 181 CONSULTA TERM DOC
 142 CONSULTA DOC DOC
 113 CONSULTA TERM TERM
  68 CONSULTA DOC DOC DOC
  57 CONSULTA TERM TERM DOC
  53 CONSULTA TERM TERM TERM
```

```
45 CONSULTA RECONSULT
37 CONSULTA TERM DOC DOC
28 CONSULTA TERM RECONSULT
25 CONSULTA DOC DOC DOC DOC
23 CONSULTA TERM TERM TERM TERM
23 CONSULTA TERM DOC DOC DOC
23 CONSULTA DOC TERM
22 CONSULTA TERM DOC TERM DOC
20 CONSULTA DOC TERM DOC
19 CONSULTA TERM DOC TERM
18 CONSULTA RECONSULT TERM
18 CONSULTA DOC TERM TERM
17 CONSULTA RECONSULT DOC
17 CONSULTA TERM TERM TERM TERM TERM
15 CONSULTA DOC DOC DOC DOC DOC
13 CONSULTA TERM TERM TERM DOC
13 CONSULTA TERM TERM DOC TERM
13 CONSULTA DOC DOC DOC DOC DOC DOC
12 CONSULTA TERM TERM TERM TERM TERM TERM
10 CONSULTA RECONSULT RECONSULT RECONSULT
10 CONSULTA RECONSULT RECONSULT
 9 CONSULTA TERM TERM TERM TERM DOC
 9 CONSULTA TERM DOC DOC TERM
 9 CONSULTA DOC DOC TERM
```

De las 4731 sesiones registradas, el 49% (2318) contienen alguna interacción aparte de la consulta, mientras que, como muestra la primera de las secuencias más frecuentes, el 46% (2190 consultas) carecen de interacción, es decir, tras formular la consulta, el usuario no ha seleccionado ningún documento o término. La explicación a este fenómeno se encuentra en que los usuarios, bien porque estén acostumbrados a determinados entornos ofimáticos, bien porque los tiempos de espera pueden alargarse, realizan varios clicks sobre el botón de envío de la consulta. Cada uno de estos clicks se registra como una nueva consulta debido a que se realiza en tiempos distintos. A esto hay que añadir que debido a la novedad del sistema, algunas conexiones no tenían como objetivo la búsqueda de información, sino la mera exploración de un nuevo servicio en el ámbito de la universidad, o de las innovaciones que ofrece el buscador. La falta de interacción tras algunas de las consultas también debe achacarse a la falta de resultados proporcionados tanto por WTB como por Google o a que los resultados que ofrecen ambos no proporcionan ninguna expectativa de relevancia. En cualquiera de los casos, estas sesiones sin interacción no proporcionan información en términos comparativos entre WTB y Google, de forma que las estadísticas que se ofrecen en adelante se refieren a las 2318 sesiones que sí contemplan una interacción.

En la tercera secuencia más frecuente únicamente se obtiene un nuevo ranking proporcionado por la selección de un término pero no llega a explorarse documento

alguno, terminando ahí la interacción con el sistema. En estas sesiones la búsqueda no ha terminado con éxito.

La segunda y sexta secuencias más frecuentes corresponden a la exploración de documentos inmediatamente tras la consulta. En estos casos, el usuario ha explorado uno o dos documentos de la lista proporcionada por Google y ha terminado en ese punto su búsqueda de información. No es posible concluir en cuántas ocasiones los usuarios han encontrado la información que buscaban, pero es obvio que no han necesitado explorar los términos sugeridos por el sistema. Es interesante, pues, estudiar las consultas asociadas a estas sesiones, intentando reconocer alguna característica común entre ellas.

El Anexo I muestra las consultas asociadas a las sesiones que empiezan y terminan con la exploración de un solo documento tras la consulta. Resulta imposible extraer alguna característica común en cuanto a su forma. Por ejemplo, no se puede concluir que sean consultas más concretas que en otros casos, aunque resulta probable que los objetivos de información de estos usuarios fueran precisos. Únicamente se observa que el porcentaje de consultas con una sola palabra es más elevado que en el caso de las sesiones que exploran un término (ver Anexo II).

La cuarta secuencia más frecuente no aporta mucha información a la evaluación del sistema. Se trata de consultas vacías que llegan a ocurrir en un 4,7% de las sesiones.

La quinta secuencia más frecuente es la que explora un término tras la consulta y a continuación selecciona un documento para su exploración, entre la lista de documentos ofrecidos por WTB y que contienen dicho término. Tampoco es posible determinar en cuántas ocasiones el usuario ha encontrado la información que busca, pero es un caso en el que el usuario que no ha hecho uso de la lista ofrecida directamente por Google. Así pues, éste también es un caso interesante para estudiar las consultas asociadas y tratar de encontrar alguna característica común a todas ellas.

El Anexo II muestra la lista de consultas asociadas a esta secuencia, así como el término seleccionado para obtener la lista sobre la que se ha seleccionado un documento. Una vez más, no es posible concluir que se tratan de consultas más o menos precisas, aunque el término seleccionado sí ofrece una pista sobre si los objetivos de búsqueda eran precisos o imprecisos. Es decir, el grado de precisión de los objetivos de búsqueda no parece relacionado con la forma de la consulta.

6.2.7 Características de los términos seleccionados

Curiosamente, existen casos en los que el usuario selecciona un término que coincide con su consulta y en prácticamente la totalidad de estos casos se trata de sintagmas. Esto muestra el interés de los sintagmas como elementos descriptivos de la información. En muy pocas ocasiones el término seleccionado no se corresponde con el objetivo de búsqueda que en principio sugiere la consulta inicial. Por ejemplo:

Consulta	**Término seleccionado**
cultivo de tejidos vegetales	introducción de cultivos nuevos

En la mayoría de los casos el término seleccionado ayuda a precisar la búsqueda o proporciona sintagmas ligeramente diferentes pero que sí se encuentran en la colección. Por ejemplo:

Consulta	**Término seleccionado**
Diseño material multimedia	diseño y elaboración de materiales multimedia
Becas	carta de solicitud de beca
calificaciones_programacion	calificación del ejercicio de programación
carrera de historia	asignaturas de la carrera de geografía e historia de la uned
centro asociado de las rozas	acceder al centro asociado de la rozas
Cine	ciclo de cine uned
compilador modula 2	compiladores de módula
COMPILADOR	compilador de pascal
concepto de simetria molecular	aprehensión de los conceptos de operación y elementos de simetría molecular
CONVALIDACIÓN	convalidación de estudios
correcion del examen de programacion i	exámen de la segunda semana de programación iii
curso de lengua catalana	cursos de idiomas
cursos alimentación	curso de alimentación y salud
cursos de catalan gratis	ejercicio de redacción en catalán
derecho penal	derecho penal ii
Deterioro del planeta	desajuste entre recursos y población en el planeta
DICCIONARIO LENGUA ESPAÑOLA	diccionarios de la lengua española e inglesa
Dietas	dietas para los profesores
Diodos	tipos de diodos
doctorados	cursos de doctorado
Electrónica Analógica	circuitos electrónicos analógicos

Obsérvese, también, que hay consultas y expresiones de la colección con errores ortográficos y falta de acentos.

6.2.8 Uso de las acciones disponibles

Como muestra la *Tabla 6-1,* en un entorno real los usuarios no han desestimado la información terminológica sino que la utilizan. *EXPLORAR TÉRMINO* es una acción presente en el 65% de las sesiones (1500 de las 2318 sesiones con interacción), lo cual supone un porcentaje elevado teniendo en cuenta que *EXPLORAR DOCUMENTO* es una acción presente en el 75% de las sesiones (1730 de las 2318 sesiones con interacción). Curiosamente, la opción de RECONSULTAR CON TÉRMINO que en un principio parecía muy prometedora tiene un uso muy restringido, apenas se ha utilizado en el 16% de las sesiones (371 de 2318). Esto puede deberse a que su utilización resulta poco intuitiva para el usuario, o a que al encontrar un concepto que satisface su búsqueda, no necesita volver a consultar al sistema, sino simplemente acceder a los documentos que lo contienen.

6.2.9 Primeras acciones de la sesión

Las evidencias más relevantes sobre la utilidad del sistema provienen de la comparación del uso de términos en relación al ranking de documentos proporcionado en primera instancia por Google. Como muestra la *Tabla 6-2, EXPLORAR TÉRMINO* es la primera acción tras la consulta en el 51% de las sesiones, mientras que *EXPLORAR DOCUMENTO* (documento dado por Google) es la primera acción en el 42%. Estos porcentajes se hacen más significativos si se consideran únicamente las consultas de más de una palabra y que, por tanto, recuperan sintagmas relacionados con la consulta. En este caso, EXPLORAR TÉRMINO es la primera acción en el 55% de las sesiones, mientras EXPLORAR DOCUMENTO es la primera acción en el 38% de las sesiones.

		% sobre todas las sesiones (2318)	% sobre sesiones con consultas de 1 palabra (886)	% sobre sesiones con consultas de más de 1 palabra (1432)
Primera ACCIÓN tras CONSULTA	EXPLORAR DOCUMENTO	42%	47%	39%
	EXPLORAR TÉRMINO	51%	45%	55%
	RECONSULTAR CON TÉRMINO	7%	8%	6%

Tabla 6-2. Primeras acciones tras la consulta

Esto significa que los términos propuestos por WTB proporcionan mayores expectativas de relevancia que el ranking de Google. Estas expectativas muestran el poder de los sintagmas para señalar información de interés, expectativas que no siempre se ven satisfechas. La *Figura 6-4* muestra la evolución de la población a lo largo del tiempo, y cómo el conocimiento que van adquiriendo los usuarios sobre el sistema termina por ir ajustando estas expectativas acercando un poco los porcentajes.

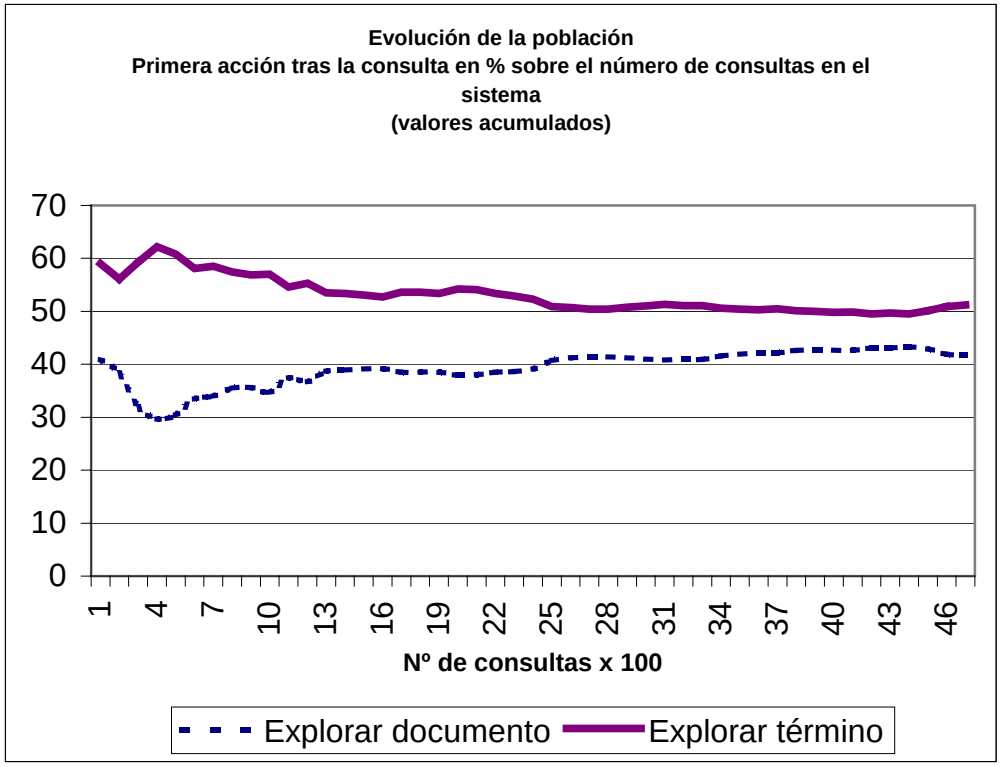


Figura 6-4. Evolución de la población respecto a la primera acción tras la consulta.

6.2.10 Últimas acciones de la sesión

Para evaluar en qué grado los términos propuestos por WTB son finalmente útiles para encontrar la información que se busca, es necesario observar cuáles han sido las últimas acciones antes de terminar la sesión. Si observamos las sesiones que terminan con la exploración de un documento y comprobamos cuál ha sido la acción anterior que ha proporcionado la lista de documentos final, tendremos una estimación, por comparación, del grado en que los términos han resultado útiles.

		% sobre todas las sesiones (2318)	% sobre sesiones con consultas de 1 palabra (886)	% sobre sesiones con consultas de más de 1 palabra (1432)
Última ACCIÓN	EXPLORAR DOCUMENTO	62%	64%	60%
	EXPLORAR TÉRMINO	31%	28%	34%
	RECONSULTAR CON TÉRMINO	7%	8%	6%

Tabla 6-3. Últimas acciones de la sesión.

La *Tabla 6-3* muestra que únicamente el 62% de las sesiones (1429) terminan con la exploración de un documento. La *Tabla 6-4* muestra cuál es la última acción antes de finalizar la sesión con la exploración de documentos.

		% sobre todas las sesiones que terminan con EXPLORAR DOC (1429)	% sobre sesiones con consultas de 1 palabra (567)	% sobre sesiones con consultas de más de 1 palabra (862)
Última ACCIÓN antes de terminar la sesión con EXPLORAR DOCUMENTO	CONSULTA	50%	57%	46%
	EXPLORAR TÉRMINO	44%	38%	47%
	RECONSULTAR CON TÉRMINO	6%	5%	7%

Tabla 6-4. Ultimas acciones antes de terminar la sesión explorando un documento.

En el 43% de estas sesiones, la última acción que ha llevado a la lista de documentos final ha sido la exploración de un término, mientras que en el 50% de los casos, la última acción fue la consulta inicial que llevó directamente al ranking de Google. Sin embargo, si consideramos únicamente las consultas que tienen más de una palabra (para las que el sistema puede ofrecer sintagmas relevantes), el porcentaje de sesiones que termina con la exploración de un documento a partir de la lista dada por WTB sube al 47% mientras que el porcentaje de las sesiones que terminan con la exploración de un documento ofrecido por Google baja al 45.6%. Es decir, para consultas de más de una palabra los porcentajes se igualan, inclinándose ligeramente hacia el uso de los términos.

El uso frecuente de los términos supone una evidencia de que la información terminológica que proporciona WTB complementa sustancialmente el ranking de documentos tradicional proporcionado por los buscadores.

La *Figura 6-5* muestra que estos resultados han ido ajustándose desde la implantación del sistema hasta estabilizarse en los datos ofrecidos.

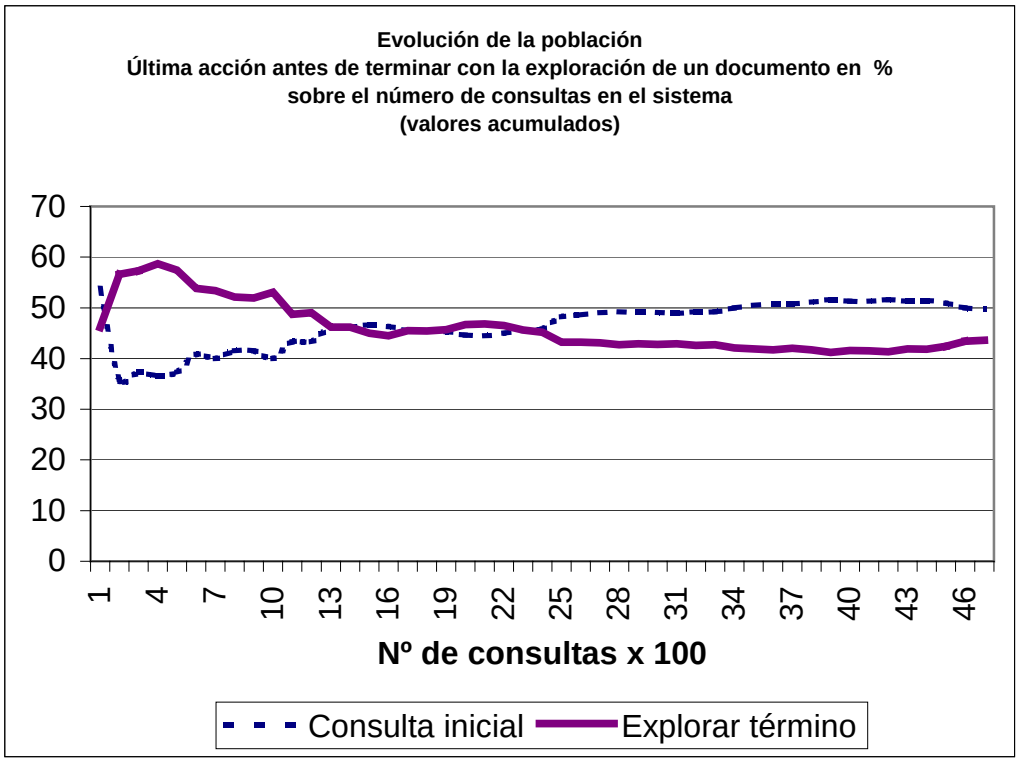


Figura 6-5. Evolución de la población respecto a la última acción antes de terminar la sesión con la exploración de un documento.

6.3 Evaluación de la recuperación translingüe de terminología

Otro de los aspectos que es necesario evaluar es la capacidad del sistema para recuperar términos relevantes en otros idiomas. Este aspecto del sistema, sin embargo, es difícil de evaluar sobre el dominio UNED ya que no cubre satisfactoriamente todos los idiomas que considera WTB. Para realizar esta evaluación, por tanto, se ha utilizado el tesauro de European Schools Treasury Browser y el prototipo cuarto.

En la versión del tesauro disponible para la evaluación (versión alfa), éste tiene 1051 descriptores con sus correspondientes traducciones para cada uno de los cuatro idiomas del tesauro (inglés, español, francés e italiano). Aproximadamente la mitad de estos términos son sintagmas de forma que pueden ser utilizados para

evaluar el sistema WTB. Los términos mono-léxicos del tesauro permiten evaluar la cobertura de los recursos que utiliza WTB para la expansión y traducción de las palabras de la consulta.

La evaluación se ha realizado de la siguiente manera. Cada uno de los descriptores en los cuatro idiomas del tesauro se ha utilizado como consulta a WTB para recuperar términos relacionados en cualquier idioma.

6.3.1 Evaluación cualitativa

La *Figura 6-6* muestra el interfaz para evaluar cualitativamente el proceso. En la primera fila se observan los términos del tesauro *therapy, terapia, thérapie y terapia* que corresponden al mismo concepto en el tesauro pero en idiomas diferentes. Todos estos términos, que son los preferidos en el tesauro, se han utilizado como consulta a WTB (primera columna). El resultado se muestra por filas en la tabla.

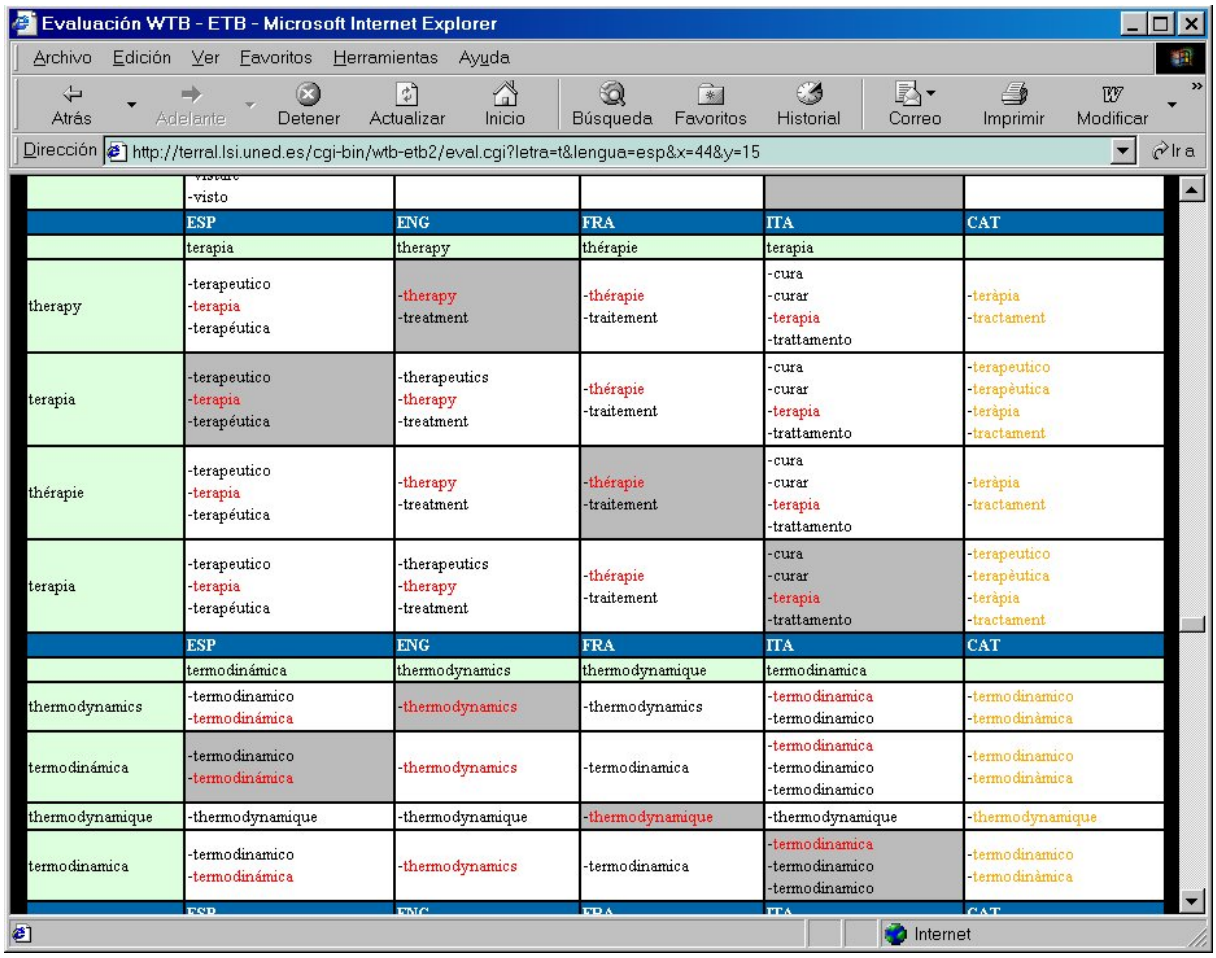


Figura 6-6. Interfaz para la evaluación cualitativa de la recuperación translingüe de terminología (términos monoléxicos)

Por ejemplo, *therapy* (en inglés), en la primera columna, recupera *terapeutico, terapia y terapéutica*, en español (misma fila, segunda columna); también recupera *therapy y treatment* en inglés (misma fila, tercera columna); etc.

6.3.2 Evaluación cuantitativa

Si entre los términos recuperados para un idioma se encuentra el término preferido en el tesauro, entonces se cuenta como un término recuperado correctamente. De esta manera es posible definir medidas de cobertura y precisión de la recuperación de terminología:

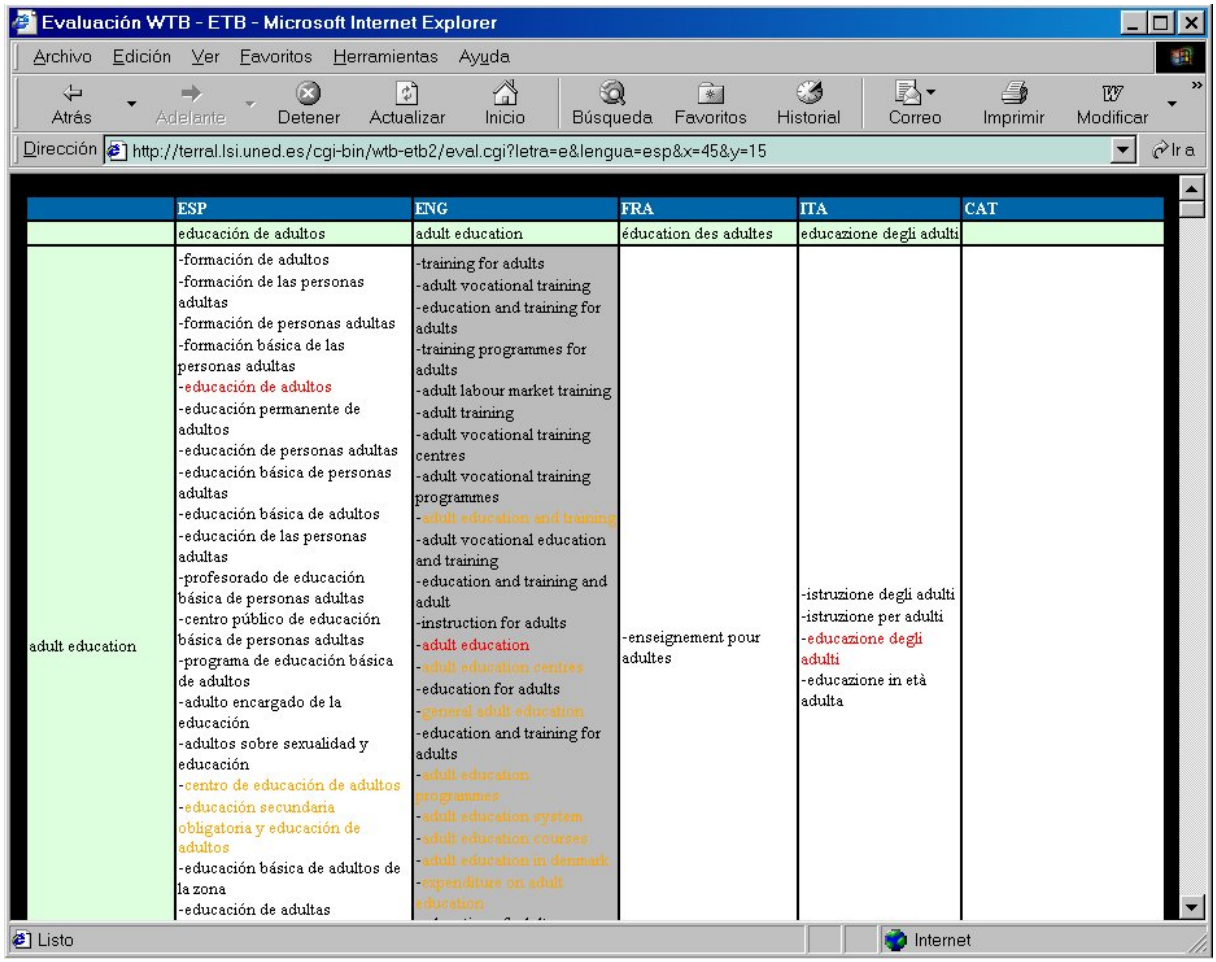
- *Cobertura*: número de términos recuperados correctamente (descriptores del tesauro) dividido por el número de términos del tesauro.
- *Precisión*: número de términos recuperados correctamente (descriptores del tesauro) dividido por el número de términos recuperados.

Observese que puede haber términos correctamente recuperados no contemplados como descritores en el tesauro. Esto quiere decir que los valores de cobertura y precisión definidos para la recuperación de terminología van a suponer una cota inferior del comportamiento real del sistema.

Por ejemplo, la *Figura 6-7* muestra cómo entre los términos recuperados *por "adult education"* en inglés, sólo el término español *"educación de adultos"* se ajusta al término preferido en el tesauro pero que, sin embargo, hay variaciones morfosintácticas *("educación de adultas", "educación de los adultos")* y semánticas *("formación de adultos")*, así como otros términos relacionados *("formación básica de las personas adultas")* que son sintagmas recuperados correctamente y que no se van a contar como tales. El ejemplo muestra que al contarse únicamente los términos coincidentes con los preferidos en el tesauro, los valores de cobertura y precisión de la recuperación de terminología obtenidos en la evaluación van a suponer una cota inferior de los que realmente obtiene el sistema.

Para ajustar un poco más el índice de cobertura no es factible contar más casos correctos sin distorsionar los resultados de forma poco predecible. Sin embargo, sí es posible ajustar un poco más la cota inferior del índice de precisión, sumando al número de sintagmas correctamente recuperados aquellos que son ajustes parciales, bien porque son sintagmas que incluyen al sintagma preferido (supersintagmas, e.g. *"centro de educación de adultos"* incluye a *"educación de adultos"*), bien porque es un sintagma incluido en el sintagma preferido por el

tesauro (subsintagmas, *"historia del siglo"* está incluido en *"historia del siglo XX")*. Aún así, siguen sin contarse las variaciones semánticas y morfosintácticas válidas. Por tanto, tanto los datos de precisión como los de cobertura de la recuperación de terminología seguirán siendo una cota inferior.



	ESP	ENG	FRA	ITA	CAT
	educación de adultos	adult education	éducation des adultes	educazione degli adulti	
adult education	-formación de adultos -formación de las personas adultas -formación de personas adultas -formación básica de las personas adultas -educación de adultos -educación permanente de adultos -educación de personas adultas -educación básica de personas adultas -educación básica de adultos -educación de las personas adultas -profesorado de educación básica de personas adultas -centro público de educación básica de personas adultas -programa de educación básica de adultos -adulto encargado de la educación -adultos sobre sexualidad y educación -centro de educación de adultos -educación secundaria obligatoria y educación de adultos -educación básica de adultos de la zona -educación de adultas	-training for adults -adult vocational training -education and training for adults -training programmes for adults -adult labour market training -adult training -adult vocational training centres -adult vocational training programmes -adult education and training -adult vocational education and training -education and training and adult -instruction for adults -adult education -adult education centres -education for adults -general adult education -education and training for adults -adult education programmes -adult education system -adult education courses -adult education in denmark -expenditure on adult education	-enseignement pour adultes	-istruzione degli adulti -istruzione per adulti -educazione degli adulti -educazione in età adulta	

Figura 6-7. Interfaz para la evaluación cualitativa de la recuperación translingüe de terminología (términos poli-léxicos)

A la hora de evaluar la recuperación translingüe de términos, es necesario recordar que los términos recuperados (lemas y sintagmas) se han extraído previamente de la colección y que, por tanto, la cobertura del sistema va a depender de la cobertura de la propia colección respecto al dominio del tesauro. Si bien la colección utilizada para esta evaluación es muy cercana al dominio del tesauro (colección multilingüe de recursos educativos), no se puede garantizar que ésta contenga todos los términos del tesauro en todos los idiomas. De hecho, los descriptores de un tesauro son expresiones de conceptos abstractos que no tienen por qué aparecer en el contenido de los documentos que se indexan. La *Tabla 6-5* muestra la cobertura de

los descriptores del tesauro en la colección, considerando apariciones exactas (incluídos acentos).

Cobertura	Español	Inglés	Francés	Italiano
Descriptores mono-léxicos en la colección	84,3%	81,9%	82,3%	81,1%
Descriptores poli-léxicos en la colección	56,5%	57,5%	54,2%	42,6%

Tabla 6-5. Descriptores del tesauro presentes en la coleción de prueba.

6.3.3 Recuperación de términos mono-léxicos

La expansión y traducción de términos mono-léxicos únicamente depende de los recursos léxicos que se utilicen. Por esta razón, la capacidad de recuperar términos mono-léxicos puede evaluarse con independencia de la colección, contando los descriptores mono-léxicos del tesauro presentes en los recursos léxicos utilizados. Esta comparación proporciona una idea de la cobertura que tienen los recursos en el dominio. La *Tabla 6-6* muestra la presencia de descriptores del tesauro en los recursos léxicos (caso monolingüe, en la diagonal), y su capacidad para recuperar términos mono-léxicos en otros idiomas de forma translingüe. La primera columna corresponde a los idiomas de partida y la primera fila a los idiomas meta. Los valores de cada casilla coresponden al porcentaje de descriptores mono-léxicos del tesauro en el idioma destino recuperados a partir de los descriptores en el idioma de partida.

Cobertura	Español	Inglés	Francés	Italiano
Español	91.66%	83.73%	60.91%	64.28%
Inglés	80.39%	97.25%	63.92%	63.92%
Francés	66.34%	61.83%	85.51%	55.96%
Italiano	67.89%	62.25%	53.89%	96.69%

Tabla 6-6. Cobertura potencial en la recuperación de descriptores mono-léxicos del tesauro de acuerdo con los recursos léxicos utilizados por WTB

La tabla muestra que la cobertura para los pares de lenguas inglés-español es significativamente más elevada que para el resto de idiomas. La razón es que tanto el inglés como el español se han visto reforzados con el uso de diccionarios bilingües además de EuroWordNet, mientras que el francés y el italiano únicamente utilizan las jerarquías de EuroWordNet.

Observese que los casos monolingües presentan un buen comportamiento por lo que la pérdida de cobertura se debe a una carencia en las relaciones de EuroWordNet entre las jerarquías de idiomas diferentes.

Como la recuperación translingüe de términos poli-lexicos depende de los lemas contenidos en los recursos disponibles, es de esperar un comportamiento de WTB mucho más pobre en la recuperación translingüe de terminología (lemas y sintagmas) que implique francés e italiano.

6.3.4 Recuperación de términos poli-léxicos

La recuperación de términos poli-léxicos depende de los sintagmas que previamente se hayan extraído de la colección y, por tanto, de que los descriptores del tesauro aparezcan en la colección. La cobertura de los descriptores en la colección para el caso monolingüe (*Tabla 6-5,* última fila), supone la cota superior de cobertura para los casos translingües. De acuerdo con esta cota superior, la *Tabla 6-7* muestra la cobertura de WTB en la recuperación de términos poli-léxicos para cada par de lenguas en porcentaje sobre la cobertura de la colección en el idioma meta.

Cobertura	Español	Inglés	Francés	Italiano
Español	63,1%	45,8%	19,9%	16,3%
Inglés	40,2%	66,5%	14,7%	7,4%
Francés	12,5%	15,6%	40,3%	7,8%
Italiano	17,1%	17,2%	8,9%	39,3%

Tabla 6-7. Cobertura de WTB en la recuperación translingüe de términos poli-léxicos en porcentaje respecto a la cobertura de la colección

Como muestra la tabla, la recuperación que implica el par de lenguas inglés-español muestra un mejor comportamiento que para el resto de idomas. La razón es que la recuperación de términos poli-léxicos se fundamenta en la combinación de términos mono-léxicos que, como ya se ha discutido, depende de los recursos léxicos utilizados. De nuevo, únicamente en el caso de los pares inglés-español EuroWordNet se complementa con dicionarios bilingües y, por esta razón, la recuperación de términos tanto monolingües como translingües presenta el mejor comportamiento con estas lenguas.

Sin embargo, estas diferencias no se corresponden con la cobertura en el caso de recuperación de términos mono-léxicos. Las diferencias de cobertura para los pares inglés-español son proporcionalmente más acusadas en el caso de términos poli-léxicos que en el de términos mono-léxicos. La explicación, una vez más, se

encuentra en las características de los recursos léxicos utilizados. Los términos mono-léxicos se corresponden casi en su totalidad con nombres comunes cuyas jerarquías en EuroWordNet están bastante bien cubiertas. Sin embargo, los términos poli-léxicos suelen implicar adjetivos cuyas jerarquías en EuroWordNet son mucho más pobres. El uso de dicionarios bilingües en los casos inglés-español, suple estas deficiencias en cuanto a adjetivos, acentuando las diferencias de cobertura para estas lenguas con respecto al resto.

Otro aspecto reseñable es que el comportamiento del español como lengua de partida es superior al del resto de idiomas. La razón se encuentra en que para el español, WTB dispone de un analizador morfológico que da todas las formas base posibles para las palabras de la consulta. La consideración de todas las formas base evita que los errores de desambiguación de categoría gramatical y lematización de la consulta afecten a la recuperación. Para el resto de lenguas, WTB no dispone de este recurso.

6.3.5 Pérdida de cobertura

Los datos anteriores evidencian la dependencia del sistema respecto a la disposición de buenos recursos léxicos y explican parcialmente la pérdida de cobertura en los casos de recuperación translingüe de terminología. Sin embargo, los valores de cobertura para los casos monolingües tampoco son elevados y esto exige un estudio adicional de cómo afecta el procesamiento lingüístico a la recuperación de terminología.

Debido a que la recuperación de términos poli-léxicos depende de los sintagmas extraídos previamente de la colección, la calidad de la recuperación no sólo depende de la colección y de la calidad de los recursos utilizados, sino también de los procesos en sí de extracción, indexación y recuperación de sintagmas. La *Tabla 6-8* muestra la pérdida de cobertura causada por los procesos de extración, indexación y selección de sintagmas. Las causas de esta pérdida son las siguientes:

1. *Proceso de extracción de sintagmas.* Como el proceso de extracción de sintagmas se fundamenta en el ajuste de patrones morfosintácticos, existen dos causas para perder sintagmas durante el proceso de extracción:
    a. *Poca exhaustividad de los patrones morfosintácticos.* Los patrones morfosintácticos son muy generales pero, aún así, existen descriptores en el tesauro que no se ajustan a ellos. Por ejemplo, el 9,6% de los descriptores poli-léxicos en francés contienen un apóstrofe (v.g. *"traitement de l'information"*). En español, el 2% de los descriptores poli-léxicos del tesauro contienen un guión (v.g. *"lengua anglo-*

americana" o *"relación padres-niño"*). Estos descriptores no encajan con los patrones morfo-sintácticos por lo que no se pueden extraer. Aparte de los sintagmas que contienen guiones, sólo hay un término español en el tesauro que no se ajusta con ningún patrón, en este caso por contener la conjunción subordinada *como*: *"idioma nacional como segundo idioma"*. Puede observarse que la pérdida de cobertura para el español es mínima en el proceso de extracción de sintagmas por lo que se puede concluir que el patrón morfosintáctico utilizado abarca muy bien las expresiones terminológicas en español.

b. Etiquetado morfosintáctico incorrecto. La otra causa para que sintagmas presentes en la colección no lleguen a extraerse es que el etiquetado morfosintáctico haya sido incorrecto. A este respecto, es destacable que con el etiquetado heurístico propuesto en este trabajo para el español, la pérdida de sintagmas en el proceso de extracción es también muy reducida.

La pérdida de cobertura debido al proceso de extracción de sintagmas oscila entre el 2.8% del español y el 17,3% del francés. Mientras que el comportamiento para el caso del español es muy bueno, los datos muestran que es necesario revisar los patrones y las herramientas de etiquetado morfosintáctico para el resto de lenguas.

2. *Proceso de selección de sintagmas.* Para mejorar la precisión en los sintagmas recuperados y ofrecidos al usuario se optó por descartar aquellos sintagmas que sólo aparecieran en un documento (df=1). Sin embargo, la evaluación muestra que de esta manera son muchos los sintagmas terminológicos que se han perdido. La pérdida de cobertura respecto a los descriptores del tesauro oscila entre el 12.9% para el caso del español y el 36.7% para el caso del italiano. Esto supone unos porcentajes demasiado elevados concluyendo que no resulta conveniente realizar este tipo de selección.

3. *Proceso de indexación y recuperación de sintagmas.* Durante la indexación y recuperación de sintagmas también se produce una pérdida de cobertura. Las causas que impiden recuperar un sintagma que ha sido previamente extraído son:

   a. Pobre expansión y traducción de las palabras de la consulta. Este problema se debe a la calidad de los recursos léxicos utilizados y ya se ha discutido en el apartado anterior
   b. Problemas de acentuación. En muchas ocasiones, la terminología aparece en las coleciones pero con errores ortográficos como la falta de acentos. En estos casos, es difícil cuantificar que porcentaje de

cobertura se ha perdido. La razón es que WTB no necesita separar documentos en idiomas diferentes y eliminar acentos llevaría a una confusión entre terminologías de diferentes idiomas. Por ejemplo, muchos términos en inglés coincidirían con términos en francés a los que se eliminaran acentos, y lo mismo ocurriría entre el italiano y el español.

c. Incorrecta lematización:
    - de las palabras componentes de los sintagmas extraídos.
    - de las palabras de la consulta.

Puede observarse que la lengua menos afectada por estos problemas es, lógicamente, el inglés (sólo un 2% de pérdida). En el caso del español, el problema de una lematización incorrecta se suple parcialmente gracias a la consideración de todas los lemas que proporciona el analisis morfológico de las palabras de la consulta. En el caso del francés y el italiano no se dispone de esta herramienta y se observan pérdidas mayores de cobertura (hasta el 34,8% en el caso del francés.

Descriptores poli-léxicos del tesauro	Español	Inglés	Francés	Italiano
Presentes en la coleción	56.5%	57.5%	54.2%	42.6%
Presentes entre los sintagmas extraídos (Pérdida en la extración de sintagmas)	54.9% (-2.8%)	50.1% (-12.9%)	44.8% (-17.3%)	40.0% (-6.1%)
Recuperados por WTB (Pérdida acumulada) (Pérdida en la indexación y recuperación de sintagmas)	40.9% (-27.6%) (-25.5%)	49.1% (-14.6%) (-2%)	29.2% (-46.1%) (-34.8%)	26.4% (-38%) (-34%)
Recuperados por WTB tras descartar df=1 (Pérdida acumulada) (Pérdida en la selección de sintagmas)	35.6% (-36.9%) (-12.9%)	38.2% (-33.5%) (-22.1%)	21.8% (-59.7%) (-25.3%)	16.7% (-60.7%) (-36.7%)

Tabla 6-8. Pérdida de cobertura en la recuperación de términos poli-lexicos

6.3.6 Precisión

En cuanto a la precisión, la *Tabla 6-10* muestra que, de media, se encuentra un término relevante de cada diez en el peor caso y de cada tres en el mejor (siempre

como cota inferior). Esto es un buen dato teniendo en cuenta que resulta muy fácil discriminar sintagmas y que WTB los agrupa y organiza jerárquicamente.

Precisión	Español	Inglés	Francés	Italiano
Español	30.69%	25.93%	30.82%	26.64%
Inglés	32.28%	37.49%	36.22%	30.12%
Francés	24.96%	25.15%	48.55%	27.82%
Italiano	28.77%	27.89%	32.47%	47.06%

Tabla 6-9. Cota inferior de precisión de WTB en la recuperación translingüe de términos mono-léxicos

Precisión	Español	Inglés	Francés	Italiano
Español	16.90%	15.30%	14.77%	12.95%
Inglés	16.88%	23.47%	12.64%	8.44%
Francés	10.96%	10.35%	22.66%	10.77%
Italiano	12.54%	10.57%	10.37%	30.24%

Tabla 6-10. Cota inferior de precisión de WTB en la recuperación translingüe de términos poli-léxicos

A este respecto, se ha evaluado la cobertura (*recall*) del sistema considerando únicamente los sintagmas que proporciona WTB como primer nivel de la jerarquía. La *Tabla 6-11, Tabla 6-12, Tabla 6-13 y Tabla 6-14* muestran el comportamiento del sistema a la hora de organizar jerárquicamente los sintagmas recuperados.

esp	esp	eng	fra	ita
recall	35.64%	26.32%	10.78%	6.946%
recall (top level)	26.50%	19.01%	7.678%	6.581%
recall (top level) % recall	74.35%	72.22%	71.18%	94.73%

Tabla 6-11. Recuperación de términos poli-léxicos en el primer nivel de la jerarquía (partiendo del español)

eng	eng	esp	fra	ita
recall	38.26%	22.73%	7.948%	3.142%
recall (top level)	28.46%	16.63%	6.469%	2.587%
recall (top level) % recall	74.39%	73.17%	81.39%	82.35%

Tabla 6-12. Recuperación de términos poli-léxicos en el primer nivel de la jerarquía (partiendo del inglés)

fra	fra	esp	eng	ita
recall	21.85%	7.037%	9.074%	3.333%
recall (top level)	16.29%	4.629%	7.037%	3.148%
recall (top level) % recall	74.57%	65.78%	77.55%	94.44%

Tabla 6-13. Recuperación de términos poli-léxicos en el primer nivel de la jerarquía (partiendo del partiendo del francés)

ita	ita	esp	eng	fra
recall	16.75%	9.683%	9.869%	4.841%
recall (top level)	14.71%	7.821%	7.821%	3.165%
recall (top level) % recall	87.77%	80.76%	79.24%	65.38%

Tabla 6-14. Recuperación de términos poli-léxicos en el primer nivel de la jerarquía (partiendo del italiano)

Puede observarse que, en general, más del 70% de los términos recuperados correctamente, se ofrecen en el primer nivel de la jerarquía. Esto supone un porcentaje suficientemente alto como para que la discriminación de sintagmas por parte del usuario sea realmente efectiva y rápida.

La evaluación muestra la dependencia del sistema respecto a la disponibilidad de recursos y herramientas de procesamiento lingüístico de calidad y muestran que cuando se dispone de ellos el comportamiento del sistema es bueno en términos de cobertura teniendo en cuenta que se ha establecido una cota inferior (puesto que no se han contado las variaciones morfosintácticas y semánticas válidas), y que si no se descartan los sintagmas qué sólo aparecen en un documento la cobertura sube entre un 5% y un 10% en términos absolutos (para la colección y tesauro utilizados).

6.4 Selección translingüe de documentos

El proceso de extracción y selección de sintagmas de WTB se ha utilizado con buenos resultados en otra tarea de evaluación relacionada con el acceso translingüe a la información: la descripción e identificación de documentos relevantes en otros idiomas. El objeto de esta tarea es proporcionar al usuario información en su propio idioma para ayudarle a identificar documentos de su interés en otros idiomas independientemente del grado de conocimiento que tenga de la lengua.

Con este fin, el grupo UNED de Procesamiento de Lenguaje Natural participó en el apartado interactivo del CLEF 2001 (iCLEF) (López-Ostenero 2001). Los participantes en este apartado evaluaron la utilidad de sus sistemas para seleccionar documentos en otro idioma. Para la evaluación se tomaron una serie de consultas (topics) y para cada una de ellas se preparó un conjunto de documentos a partir de las recuperaciones efectuadas por los sistemas CLIR participantes en la edición de CLEF del año 2000.

Los sistemas participantes en iCLEF debían proporcionar una descripción de cada uno de los documentos para que un grupo de usuarios seleccionara los que a su juicio fueran relevantes de acuerdo con esa descripción. El tiempo para seleccionar o descartar los documentos fue acotado para cada consulta. De esta forma se obtuvieron estadísticas de acierto, error, precisión, número de documentos explorados, etc.

La evaluación de cada sistema se realizó por comparación con los resultados obtenidos cuando la descripción de los documentos era resultado de una traducción automática del documento mediante uno de los mejores sistemas de traducción automática del mercado (Systran professional 3.0).

La descripción de los documentos que proporcionaba el sistema presentado por el grupo UNED de PLN se basa en la extracción de los sintagmas del documento mediante el software de WTB, y su alineación con sintagmas extraídos de toda la colección también con WTB en el idioma destino. Al usuario se le presentaban los sintagmas alineados total o parcialmente como descripción del documento.

El sistema de la UNED obtuvo el mejor resultado entre los participantes y mejoró los resultados obtenidos con el sistema de referencia (traducción automática mediante Systran) (López-Ostenero 2001).

6.5 Otras tareas de aplicación y evaluación

Como sugiere la evaluación anterior, las relaciones que proporciona WTB entre terminología y documentos no sólo permite llevar a cabo tareas de acceso a la información, sino también tareas relacionadas con la obtención de terminología y la utilización de vocabularios controlados en la organización y recuperación de recursos.

6.5.1 Identificación de terminología

Como muestran la *Figura 6-6 y Figura 6-7* en el apartado anterior, el sistema permite obtener términos que pueden enriquecer un tesauro con nuevos sinónimos, además de ofrecer una perspectiva de los términos que realmente aparecen en los documentos reales. Esto supone una ayuda para el trabajo de documentalistas en la construcción de vocabularios.

Evidentemente, para tareas de identificación de nueva terminología, al sistema no sólo se le puede consultar con términos ya contemplados en un vocabulario, sino que se puede recuperar terminología a partir de una cuantas palabras clave elegidas por un documentalista. De esta forma no sólo se pueden enriquecer tesauros con términos candidatos sino que, en una etapa anterior, también se pueden proponer los primeros candidatos a formar parte del vocabulario.

6.5.2 Vía de acceso a un tesauro

Una vez construidos los vocabularios controlados, su utilización para indexar documentos acarrea el problema de que en la recuperación el usuario debe conocer de antemano cuáles son los términos de indexación para utilizarlos en la consulta. Sin embargo, el usuario no tiene por qué conocer el vocabulario que se ha utilizado en la indexación. Para salvar la distancia entre los términos utilizados en las consultas y el vocabulario controlado utilizado en la indexación de la colección, también es posible utilizar WTB.

Para abordar el problema basta indexar por sus componentes normalizadas los términos del vocabulario controlado del mismo modo que a lo largo del trabajo se han indexado los sintagmas extraídos automáticamente de la colección. De esta forma, la expansión, traducción y combinación de las palabras de la consulta para recuperar términos del vocabulario controlado permiten salvar la distancia entre el vocabulario del usuario y el de la colección.

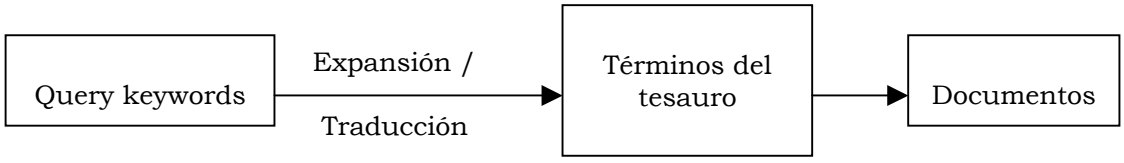


Figura 6-8. Uso de WTB como vía de acceso a un tesauro

Capítulo 7

Conclusiones

Los sistemas de acceso a la información todavía no incorporan soluciones satisfactorias a los problemas de ambigüedad léxica, variación morfosintáctica, variación semántica y variación translingüe. Si bien las técnicas automáticas de procesamiento lingüístico resultan buenas candidatas para abordar estos problemas, los experimentos llevados a cabo por diversos autores hasta la fecha no han obtenido resultados determinantes en cuanto al papel que pueden desempeñar las técnicas lingüísticas automáticas en los modelos tradicionales de recuperación de documentos.

La primera parte del trabajo ha tratado de discernir si la falta de resultados satisfactorios se debe a la falta de precisión en el procesamiento lingüístico automático o a que las técnicas no resultan apropiadas dentro del modelo tradicional de recuperación y ranking de documentos. Para ello, se han mostrado una serie de experimentos de recuperación de documentos sobre una colección etiquetada manualmente en todos los niveles léxicos. De esta forma, los resultados obtenidos han quedado libres de los errores de un procesamiento automático, permitiendo determinar si las técnicas lingüísticas (en una situación ideal) podían suponer o no estrategias adecuadas para mejorar la recuperación.

Los experimentos llevados a cabo sobre esta colección muestran que, en un modelo tradicional de recuperación y ranking de documentos, ni la desambiguación de la categoría gramatical, ni la desambiguación del sentido de las palabras producen

unas mejoras significativas en la recuperación que justifiquen el coste de procesamiento que introducen.

Respecto a la detección y distinción de compuestos léxicos, se ha propuesto un método automático para distinguir semánticamente los compuestos léxicos de WordNet y se ha estudiado cómo afecta su distinción al modelo tradicional de recuperación de documentos. Los resultados muestran que la consideración de sintagmas no mejora la recuperación salvo en el caso de que se distingan y consideren únicamente compuestos exocéntricos. Sin embargo, las consultas de la colección de test incluyen un número muy limitado de compuestos exocéntricos por lo que las diferencias en los resultados de recuperación son demasiado pequeñas.

La desambiguación del sentido de las palabras, sin embargo, abre la posibilidad de una indexación basada en *synsets* (índices de conjuntos de sinónimos) que resulta bastante prometedora. Se han estudiado los *synsets* que se pueden utilizar en una indexación conceptual sin el problema de anotación semántica *(Synsets de Variantes Monosémicas*). Además, se han reproducido y contrastado los experimentos de (Sanderson 1994) utilizando *synsets* de WordNet en lugar de pseudo-palabras. Los experimentos que se han llevado a cabo sobre la colección de prueba IR-SEMCOR concluyen que la indexación con *synsets* es más beneficiosa y resistente a errores de lo que cabía deducir a partir de los experimentos de Sanderson.

Estos resultados llevaron al *Grupo de Procesamiento del Lenguaje Natural de la UNED* a desarrollar un motor de búsqueda (*ITEM Search Engine*) basado en indexación conceptual sobre synsets que también se ha presentado en este trabajo. La ventaja adicional de una indexación conceptual basada en *synsets* es que permite de forma natural una recuperación translingüe de documentos, ya que los *synsets* de diferentes idiomas están conectados entre sí en EuroWordNet en un índice independiente de la lengua.

Sin embargo, la experiencia con este buscador ha puesto de manifiesto las dificultades prácticas de un sistema de estas características. Por una parte, la secuencia de procesamiento lingüístico (lematización, etiquetado de categoría, detección de expresiones multipalabra y desambiguación del sentido de las palabras) añade un coste de procesamiento que impide la indexación de colecciones medianas en un tiempo razonable. A esto hay que añadir que las técnicas existentes de desambiguación del sentido de las palabras tampoco son suficientemente precisas y que los conceptos de WordNets no son buenas unidades de traducción, por lo que la evaluación cualitativa del sistema no ha sido plenamente satisfactoria.

Con estos antecedentes, la segunda parte del trabajo ha explorado una nueva posibilidad de abordar los problemas de ambigüedad léxica, variación terminológica

y multilingüismo. En lugar de aplicar las técnicas lingüísticas a la indexación de información, subordinándolas al modelo clásico de recuperación de documentos, se ha explorado la posibilidad de ofrecer al usuario un nivel intermedio de información resultado de un procesamiento parcial del lenguaje natural. Este nivel de información se ha concretado en una nueva área de terminología extraída automáticamente a partir de la colección y particularizada de acuerdo con la consulta. El modelo afecta a la indexación, a la recuperación y enriquece la interacción.

El modelo de indexación propuesto incorpora una extracción automática de sintagmas con criterios terminológicos. Esta extracción se ha realizado sobre la base de patrones morfosintácticos que requieren un etiquetado previo de los textos. El etiquetado es un proceso costoso que puede resultar inviable cuando se procesan cientos de miles de documentos. Para superar esta limitación, se ha propuesto e implementado un proceso de etiquetado heurístico dirigido a la tarea concreta de extracción de sintagmas terminológicos. Este etiquetado se basa en el análisis morfológico de las palabras y en un juego de tres etiquetas para asignar una sola etiqueta a cada palabra. Esta etiqueta es siempre la misma por lo que el proceso de etiquetado se reduce a construir un lexicón sobre el que se pueda obtener el lema y etiqueta de las palabras de la colección. La evaluación ha mostrado que no sólo es un proceso más eficiente sino que la pérdida de sintagmas en el proceso de extracción se reduce notablemente (2.8%)

La extracción de sintagmas basada en el ajuste de patrones morfosintácticos proporciona una gran cantidad de sintagmas que no es necesario indexar. Para abordar esta cuestión, se ha propuesto e implementado un proceso de selección de sintagmas basado en la frecuencia de aparición de los sintagmas en la colección y a criterios de subsunción de subsintagmas.

El modelo propuesto de recuperación se basa en la traducción y expansión por sinónimos de las palabras de la consulta para recuperar sintagmas relacionados. La co-ocurrencia de palabras en un mismo sintagma presente en la colección, es una restricción muy fuerte que desambigua implícitamente categorías gramaticales, sentidos y determina los sinónimos y traducciones más adecuados. De esta manera, los sintagmas recuperados suponen variaciones morfosintácticas, semánticas y translingües de la consulta. Estos sintagmas, al ser extraídos de la colección, proporcionan acceso directo a los documentos que los contienen, convirtiéndose en una vía alternativa de acceso a la información. Los términos obtenidos a partir de la consulta se consideran también para ofrecer el ranking tradicional de documentos.

El modelo de interacción se basa en el uso de dos áreas diferentes, una que organiza y presenta al usuario los términos recuperados, y otra que presenta al

usuario una lista de documentos. Sobre este modelo se han definido una serie de acciones que proporcionan al usuario vías alternativas de acceso a la información que permiten abordar:

1. Situaciones de búsqueda en las que no se cumplen los presupuestos implícitos de una recuperación basada en la búsqueda y ordenación de documentos como, por ejemplo, que las necesidades de información se vean afectadas por los resultados de la búsqueda. En este caso, el sistema ofrece al usuario nuevas alternativas de acceso a la información a través sintagmas.
2. Situaciones de imprecisión. El sistema ayuda a precisar las necesidades de información de los usuarios sugiriendo sintagmas más específicos y sintagmas relacionados con la consulta pero con otras palabras.
3. Situaciones en las que el usuario no conoce la terminología propia del dominio. En estos casos, las variaciones terminológicas que ofrece el sistema pueden ayudarle a superar esta barrera.
4. Situaciones en las que el usuario puede entender un texto en otro idioma pero no puede expresar su consulta en ese idioma.

La propuesta se ha llevado a la práctica con el desarrollo de las diferentes versiones y prototipos de *Website Term Browser (WTB*). El desarrollo incremental del sistema ha permitido abordar los problemas técnicos de la propuesta y definir las especificaciones finales del sistema.

Las evaluaciones diseñadas para los sistemas de recuperación de documentos no son aplicables a *WTB.* Tampoco resultan apropiadas las evaluaciones diseñadas para los sistemas de búsqueda interactiva, ni para los sistemas que proporcionan interactividad al tratar los problemas de multilingüismo. Por esta razón, se ha diseñado un nuevo marco de evaluación dirigido a:

1. Comprobar si la nueva área de términos resulta de utilidad para los usuarios.
2. Evaluar la capacidad del sistema para recuperar terminología de forma translingüe.
3. Determinar su capacidad para tratar grandes volúmenes de información.
4. Estudiar la aplicación del sistema a otras tareas como la selección translingüe de documentos, la identificación de terminología en la construcción de vocabularios controlados o el acceso a los términos de un tesauro.

La utilidad del área de términos se ha evaluado comparando el uso que le dan los usuarios, frente al uso que le dan al ranking de documentos proporcionado por uno de los mejores buscadores en Internet. La evaluación se ha realizado en un entorno

real de trabajo, registrando las interacciones de usuarios con necesidades reales de información. Para ello, se han registrado, almacenado y analizado las interacciones de más de 2000 sesiones mostrando que los usuarios estiman de utilidad el área de términos y que supone un complemento al ranking tradicional de documentos.

La exploración de un término es una acción presente en el 65% de las sesiones con interacción. La primera acción tras la consulta es mayoritariamente la exploración de un término (60%) frente a la exploración directa de un documento (39%). Esto significa que los términos propuestos por WTB proporcionan mayores expectativas de relevancia que el ranking de Google. Estas expectativas muestran la capacidad de los sintagmas para señalar información de interés.

El porcentaje de sesiones que termina con la exploración de un documento a partir del ranking ofrecido por la selección de un sintagma de WTB es del 47% mientras que el porcentaje de las sesiones que terminan con la exploración de un documento ofrecido por Google es del 45.6%. Esto confirma que la información terminológica que proporciona WTB complementa sustancialmente el ranking de documentos proporcionado por los buscadores tradicionales.

La evaluación de la recuperación de terminología en el caso translingüe se ha realizado utilizando un tesauro multilingüe construido manualmente. Los datos evidencian la dependencia del sistema respecto a la disposición de recursos y herramientas de procesamiento lingüístico en cada idioma y muestran que cuando se dispone de ellos el comportamiento del sistema es bueno en términos de cobertura, aún sin considerar variaciones morfosintácticas y semánticas válidas que ofrece WTB. Sin embargo, es necesario seguir trabajando en el desarrollo de recursos y herramientas de Procesamiento de Lenguaje Natural. La evaluación ha permitido contrastar las deficiencias de recursos léxicos como EuroWordNet (deficiencia en las jerarquías de adjetivos y en las relaciones entre jerarquías de distintos idiomas), así como la conveniencia de disponer de heramientas robustas de PLN como, por ejemplo, analizadores morfológicos.

El sistema ofrece al menos un término relevante por cada tres en el mejor caso sin contar variaciones terminológicas válidas. Esto es un buen dato teniendo en cuenta que resulta muy fácil discriminar sintagmas y que WTB los agrupa y organiza jerárquicamente. Además, el 70% de los términos relevantes se ofrece en el primer nivel de la jerarquía.

En cuanto a la capacidad del sistema para tratar grandes cantidades de información, la incorporación de técnicas y algoritmos bien conocidos en el campo de Recuperación de Información han permitido procesar las colecciones de español e inglés de CLEF 2001 con aproximadamente 1 Gb de textos. Los tiempos de

indexación y de respuesta son viables, aunque las necesidades de espacio en disco requieren la incorporación de algoritmos IR más sofisticados si se quiere procesar colecciones de mayor tamaño.

Por último, se han presentado los resultados que ha obtenido un sistema de alineación de sintagmas en el apartado interactivo de CLEF 2001. Este sistema ha utilizado los sintagmas obtenidos por WTB como elementos de descripción en la tarea de selección de documentos en otro idioma. El sistema presentado obtuvo el mejor resultado entre los participantes y mejoró los resultados obtenidos con el sistema de referencia (traducción automática mediante Systran).

7.1 Líneas futuras de trabajo

En este trabajo se ha presentado un enfoque en el que el sistema ayuda al usuario a contextualizar su consulta. Para ello, se le ofrece un nivel intermedio de información resultado de realizar una serie de inferencias lingüísticas sobre la consulta y que permiten considerar variaciones morfosintácticas, semánticas y translingües presentes en la colección. Las líneas futuras de trabajo se dirigirán a:

1. Mejorar aspectos técnicos de eficiencia y efectividad del sistema para poder abordar volúmenes de información mayores en varios órdenes de magnitud. Para ello, será imprescindible incorporar algoritmos bien conocidos en el área de Recuperación de Información, dirigidos a tratar grandes cantidades de texto.

2. Utilizar información navegacional como, por ejemplo, el texto y el contexto de los enlaces asociados a las páginas web de destino. Así como el título de un documento es uno de los elementos más informativos del mismo, también el texto de los enlaces que apuntan al documento es una información indexable y relevante. Esta información supone, en muchas ocasiones, una clasificación informal del documento apuntado, permitiendo indexar un documento no por el texto que contiene, sino por los textos que lo describen y clasifican. Una vez más, la recuperación, ordenación y presentación al usuario de esta información textual permite construir un nuevo nivel interactivo de acceso a la información en la línea de *WTB*.

3. Considerar nuevas inferencias lingüísticas que enriquezcan los niveles intermedios de información que se ofrecen al usuario en la interacción.
    a. En el nivel sintáctico, debe estudiarse la forma de enriquecer el modelo considerando no sólo sintagmas nominales sino también

verbales. La identificación de estructuras *Sujeto-Acción-Objeto* resulta una extensión interesante del modelo que permitirá detectar de forma más rica fragmentos relevantes de información. Para ello, será necesario reconsiderar la utilización de las herramientas de procesamiento lingüístico y estudiar en qué medida la interacción puede ayudar a relajar el procesamiento para hacerlo computacionalmente viable.

b. Análogamente, el enriquecimiento del modelo en el nivel de discurso puede ir acompañado de inferencias lingüísticas como, por ejemplo, la resolución de algunos tipos de anáfora que permitan considerar los referentes como elementos adicionales de indexación de fragmentos textuales.
c. En el nivel pragmático, es interesante distinguir necesidades y tipos de información. Dirigir la interacción hacia la determinación de *para qué* se busca una pieza de información es otra forma de contextualizar (e interpretar) la consulta y, por tanto, de acceder a la información que se busca. Análogamente, determinar qué tipo de información está buscando un usuario es un proceso que también puede dirigir la interacción.

La idea que subyace a estas líneas de trabajo es la misma: el acceso a la información textual es un proceso de construcción de un contexto en el cual la consulta finalmente cobra el significado pretendido por el usuario. Ese contexto construido es la información que busca el usuario, ya sea un documento, un conjunto de fragmentos, un resumen automático o una síntesis. La construcción del contexto y, por tanto, del acceso a la información, se contempla como el recorrido de un camino. Un sistema puede construir los caminos posibles pero la elección final del recorrido debe realizarla el único que conoce la finalidad de la búsqueda y que puede dar significado a la información: el usuario.

Capítulo 8

Bibliografía

Agirre, E. and Rigau G. Word Sense Disambiguation using conceptual density. Proceedings of COLING'96; 1996.

Amo, P. Tesis doctoral: Análisis sintagmático automático y su aplicación a la Recuperación de Información: Escuela Politécnica, Universidad de Alcalá; 2000.

Anick, P. G. and Tipirneni S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. Proceedings of 22nd ACM SIGIR Conference Research and Development in Information Retrieval. 1999; 153-159.

Baeza-Yates, R. and Ribeiro-Neto B. Modern Information Retrieval. Addison-Wesley; 1999.

Ballesteros, L. and Croft W. B. Dictionary methods for Cross-Lingual Information Retrieval. Database and Expert Systems Applications. 1996.

Ballesteros, L. and Croft W. B. Resolving Ambiguity for Cross-Language Information Retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998; 64-71.

Bates, M. J. Where should the person stop and the information search interface

start? Information and Management. 1990; 26:575-591.

Bely, N. Borillo A. Virbel J. and Siot-Decauville N. Procédures d'analyse semantique appliquées a la documentation scientifique. Gauthier-Villars. 1970.

Bernier, C. L. and Heumann K. F. Correlative Indexes, III. Semantic relations among semantemes. The technical thesaurus. American Documentation. 1957; 8:211-220.

Boughanem, M. Chrisment C. and Nassr N. Investigation on disambiguation in CLIR aligned corpus and bidirectional translation-based strategies. Peters, C. et al. Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001; LNCS 2406. Springer-Verlag; 2002; pp. 158-168.

Bourigault, D. Surface grammatical analysis for the extraction of terminological noun phrases. Proceedings of 14th International Conference on Computational Linguistics, COLING'92. 1992; 977-981.

Brill, E. A simple rule-based part of speech tagger. Proceedings of the 3rd Conference on Applied Natural Language Processing; 1992.

Buitelaar, P. CoreLex: systematic polysemy and underspecification (Ph.D. thesis). Department of Computer Science: Brandeis University, Boston; 1998.

Callan, J. Croft B. and Harding S. The INQUERY retrieval system. Proceedings of the 3rd International Conference on Database and Expert Systems applications; 1992.

Carbonell, J. G. Yang Y. Frederking R. E. Brown R. D. Geng Y. and Lee D. Translingual Information Retrieval: A comparative evaluation. Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence. 1997.

Carmona, J. Cervell S. Màrquez L. Martí M. A. Padró L. Placer R. Rodríguez H. Taulé M. and Turmo J. An environment for morphosyntactic processing of unrestricted Spanish text. Proceedings of LREC'98. 1998.

Chugur, I. Gonzalo J. and Verdejo F. Sense distinctions in NLP applications. Proceedings of ONTOLEX'2000: Ontologies and Lexical Knowledge Bases; Sofia. 2000.

Chugur, I. Peñas A. Gonzalo J. and Verdejo F. Monolingual and bilingual dictionary approaches to the enrichment of the Spanish WordNet with adjectives. Proceedings of NAACL Workshop on WordNet and other lexical resources: applications, extensions and customizations.; Carnegie Mellon University,

Pittsburgh. 2001.

Cleverdon, C. W. The Cranfield tests on index language devices. ASLIB Proceedings. 1967; 173-192.

Croft, W. B. Turtle H. R. and Lewis D. D. The use of phrases and structured queries in information retrieval. Proceedings of 14th SIGIR Conference on Research and Development in Information Retrieval. 1991; 32-45.

Dillon, M. and Gray A. S. Fully automatic syntax-based indexing. Journal of the American Society for Information Science. 1983; 34(2):99-108.

Fagan, J. L. The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. Journal of the American Society for Information Science. 1989; 40(2):115-132.

Fernández-Amorós, D. Gonzalo J. and Verdejo F. The role of conceptual relations in Word Sense Disambiguation. Proceedings of the 6th International Workshop on Applications of Natural Language for Information Systems NLDB'2001; 2001.

---. The UNED systems at Senseval-2. Proceedings of Senseval-2. 2002.

Forsyth R., Rada R. Adding an edge in Machine Learning: applications in Expert Systems and Information Retrieval. Ellis Horwood Ltd. 1986; 198-212.

Frakes, W. B. and Baeza-Yates R. Information Retrieval. Data Structures and Algorithms. Prentice Hall PTR; 1992.

Frank, E. Paynter G. Witten I. Gutwin C. and Nevill-Manning C. Domain-specific keyphrase extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan-Kaufmann. 1999; 668-573.

Frantzi, K. T. and S. Ananiadou. The C-value/NC-value domain independent method for multiword term extraction. Journal of Natural Language Processing. 1999; 6(3):145-180.

Gonzalo, J. Chugur I. and Verdejo F. Sense Clusters for Information Retrieval: Evidence from Semcor and the InterLingual Index. Proceedings of the ACL'2000 workshop on Word Senses and Multilinguality ; Hong Kong. 2000.

Gonzalo, J. Peñas A. and Verdejo F. Lexical Ambiguity and Information Retrieval Revisited. Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC, Maryland. 1999a; 195-202.

Gonzalo, J. Verdejo F. and Chugur I. Using EuroWordNet in a concept-based approach to Cross-Language Text Retrieval. Applied Artificial Intelligence, Special Issue on Multilinguality in the Software Industry: the AI Contribution. 1999b.

Gonzalo, J. Verdejo F. Chugur I. López F. y Peñas A. Extracción de relaciones semánticas entre nombres y verbos en EuroWordNet. Revista De La Sociedad Española De Procesamiento Del Lenguaje Natural. 1998a; 23:97-103.

Gonzalo, J. Verdejo M. F. Chugur I. and Cigarrán J. Indexing with WordNet synsets can improve Text Retrieval. Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems; 1998b.

Hearst, M. A. and Pedersen J. O. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. Proceedings of 19th ACM SIGIR Conference on Research and Development in Information Retrieval. 1996.

Hull, D. and Gefenstette G. Querying across languages: A dictionary-based approach to multilingual information retrieval. Proceedings of 19th International Conference on Research and Development in Information Retrieval. 1996.

Hutchins, W. J. Languages of indexing and classification. Peter Peregrinus, Stevenage. 1975.

Jacquemin, C. Spotting and Discovering Terms through NLP. MIT Press, Cambridge MA. 2000.

Jones, S. and Staveley M. S. Phrasier: a System for Interactive Document Retrieval Using Keyphrases. Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval. 1999; 160-167.

Krovetz, R. Homonymy and polysemy in Information Retrieval. ACL/EACL'97; 1997.

Krovetz, R. and Croft W. B. Lexical ambiguity and information retrieval. ACM Transactions on Information Systems. 1992; 10(2):115-141.

Lancaster, F. W. Vocabulary control for information retrieval. Information Resources Press. 1972.

López-Ostenero, F. Gonzalo J. Peñas A. and Verdejo F. Noun phrase translations for Cross-Language Document Selection. Working Notes for the CLEF 2001 Workshop. 2001; 231-241.

Mandala, R. Tokunaga T. and Tanaka H. Combining Multiple Evidence from

Different Types of Thesaurus for Query Expansion. Proceedings of 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval. 1999; 191-197.

Márquez, L. and Padró L. A flexible POS tagger using an automatically acquired language model. Proceedings of ACL/EACL'97. 1997.

McCarley, J. S. Should we translate the documents or the queries in Cross-Language Information Retrieval? Proceedings of 37th Annual Meeting of ACL; 1999.

Miller, G. A. Leacock C. Tengi R. and Bunker R. T. Asemantic concordance. Proceedings of the ARPA Workshop on Human Language Technology. Morgan Kauffman. 1993.

Miller, G. Beckwith C. Fellbaum D. Gross D. and Miller K. Five papers on WordNet. Princeton University; 1990; CSL report 43, Cognitive Science Laboratoy.

Nevill-Manning, C. G. Witten I. H. and Paynter G. W. Lexically-generated subject hierarchies for browsing large collections. International Journal of Digital Libraries. 1999; 2(2/3):111-123.

Oard, D. A comparative study of query and document translation for Cross-Language Information Retrieval. Proceedings of 3rd Conference of the Association for Machine Translation in the Americas; 1998.

---. Evaluating Cross-Language Information Retrieval: Document selection. Cross-Language Information Retrieval and Evaluation: Proceedings of CLEF'2000: Springer-Verlag; 2001.

Ogden, W. Cowie J. Davis M. Ludovik E. Molina-Salgado H. and Shin H. Getting information from documents you cannot read: an interactive cross-language text retrieval and summarisation system. Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access: 1999.

Padró, L. A Hybrid Environment for Syntax-Semantic Tagging, Ph.D Thesis. Barcelona: Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya; 1998.

Paynter, G. W. and Witten I. H. A combined phrase and thesaurus browser for large document collections. Proceedings of the JCDL'2001 Workshop on the Technology of Browsing Applications. 2001a.

Paynter, G. W Nevill-Manning C. G. and Witten I. H. Phrase hierarchy inference. Proceedings of the JCDL'2001 Workshop on the Technology of Browsing Applications. 2001b.

Peñas, A. Chugur I. Gonzalo J. y Verdejo F. Incorporación de Adjetivos al WordNet español. Revista De La Sociedad Española De Procesamiento Del Lenguaje Natural. 2000; 27:81-88.

Pickens, J. and Croft W. B. An Exploratory analysis of Phrases in Text Retrieval. Proceedings of RIAO 2000 Conference, Paris. 2000; 1179-1195.

Pirkola, A. The effects of query structure and dictionary setups in dictionary-based Cross-Language Information Retrieval. Proceedings of 21st International Conference on Research and Development in Information Retrieval. 1998.

Richardson, R. and Smeaton A. F. Using WordNet in a knowledge-based approach to Information Retrieval. BCS-IRSG Colloquium; Crewe. 1995.

Salton, G. Automatic information organisation and retrieval. McGraw-Hill. 1968.

---. The SMART retrieval system. Prentice-Hall, Englewood Cliffs, N.J. 1971.

---. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). Journal of the American Society for Information Science. 1972; 23(2):75-84.

Salton, G. and McGill M. Introduction to Modern Information Retrieval. New York: McGraw-Hill; 1983a.

Salton, G. Fox E. A. and Wu H. Extended boolean Information Retrieval. Communications of the ACM. 1983b; 26(11):1022-1036.

Sanderson, M. Word Sense Disambiguation and Information Retrieval. Proceedings of 17th International Conference on Research and Development in Information Retrieval; 1994.

---. Retrieving with good sense. Information Retrieval. 2000; 2(1):45-65.

Sanderson, M. and Croft B. Deriving concept hierarchies from text. Proceedings of 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval. 1999; 206-212.

Schütze, H. and Pedersen J. Information Retrieval based on word senses. Fourth Annual Symposium on Document Analysis and Information Retrieval; 1995.

Smeaton, A. F. and Quigley A. Experiments on using semantic distances between words in image caption retrieval. Proceedings of the 19th International Conference on Research and Development in Information Retrieval; 1996.

Sparck Jones, K. What is the Role of NLP in Text Retrieval? Natural Language Information Retrieval, Ed. T. Strzalkowski, Kluwer Academic Publishers. 1999.

Strzalkowski, T. Natural language Processing Information Retrieval. Kluwer, Boston, MA. 1999.

Sussna, M. Word sense disambiguation for free-text indexing using a massive semantic network. Proceedings of the International Conference on Information and Knowledge Management CIKM'93. 1993; 67-74.

Ureña, A. Tesis doctoral: Resolución de la ambigüedad léxica en tareas de clasificación automática de documentos. Granada: Departamento de Lenguajes y Sistemas Informáticos, Universidad de Granada; 2000.

Verdejo, F. Gonzalo J. Peñas A. López F. and Fernández D. Evaluating wordnets in a Cross-Language Retrieval Environment: the ITEM search engine. Proceedings of the 2nd LREC. 2000; 1769-1774.

Vickery, B. C. Thesaurus - A new word in documentation. Journal of Documentation. 1960; 16:4:181-189.

Vivaldi, J. Tesis doctoral. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. Barcelona: Departament de Llenguatges i Sistemes Informàtics, Univeritat Politécnica de Catalunya; 2001.

Voorhees, E. M. Using WordNet to disambiguate word sense for text retrieval. Proceedings of ACM SIGIR Conference. 1993; 171-180.

---. Query expansion using lexical-semantic relations. Proceedings of the 17th ACM-SIGIR Conference. 1994; 61-69.

Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet. 1998.

Vossen, P. Peters W. and Gonzalo J. Towards a Universal Index of Meaning. Proceedings of ACL/SIGLEX'99: Standarizing Lexical Resources. 1999.

Wacholder, N. and Manning N. The technology of Phrase Browsing apllications. SIGIR FORUM. 2001; 35(1):18.

Wallis, P. Information retrieval based on paraphrase. Proceedings of PACLING Conference. 1993.

Witten, I. H. et al. Greenstone: a comprehensive open-source digital library software system. Proceedings of ACM Conference on Digital Libraries. 1999a; 113-121.

Witten, I. H. Moffat A. and Bell T. C. Managing Gigabytes. Compressing and Indexing Documents and Images. Second Edition ed. Morgan Kaufmann Publishers; 1999b.

Wolff, J. G. An algorithm for the segmentation of an artificial language analogue. British Journal of Psychology. 1975; 66:79-90.

---. Language acquisition and the discovery of phrase structure. Language and Speech. 1980; 23(3):255-269.

Ziv, J. and Lempel A. Compression of individual sequences via variable-rate coding. IEEE Trans Information Theory. 1978; IT-24(5):530-536.

Anexos

Anexo I: Consultas de sesiones en WTB que empiezan y terminan con la exploración de un solo documento

programacion
"!importancia de los valores"
"ANALISIS III"
"before the rain"
"curriculum profesorado"
"dos culturas"
"Edith Checa"
"Juan Antonio Rodríguez González"
"la evaluacion en la educacion primaria"
"oferta empleo UNED"
902
activo fijo
actos de habla
adquisición de la gramática
ADSL
AECA
affinsa
aguado
alfabeto griego
alhambra
alicia garcía falgueras
ambientalismo
analema
anselmo peñas
antropología social
anuncios
aprendizaje
aprobados programacion iii
apuntes
Artacho
arte conceptual
arte portugués

asignaturas de ciencias politicas
assessment
Auster
barroco
basalla
Bases de datos
bases de programación
bcl-2
Beatriz Barros
Beatriz Barros Blanco
becas para Cuba
biblioteca
bici
bid for power
cabrera
cálculo, construcción y ensayo
CALIFICACIONES
calificaciones
calificaciones derecho natural
CARREFOUR
CASTRO GIL
catalizador
celina
centro asociado pamplona
centros asociados
cervantes
CLEPSIDRA
cognitivo
coie
como saber si un hombre no ha tenido relaciones
compilador
Compilador de Modula2
Compilador Modula 2
COMPILADOR MODULA
Compilador Modula
compilador modula
compilador modula2
compiladores
concepto de grupo
conducta maternal
conferencias
Convalidaciones
convalidaciones
convocatoria
convocatoria profesorado
correos
creus
cristamine
cuadernillo
Cuadernillo de Prácticas
cultura clasica

cultura griega
curso de lectura de imagen
curso wap para telefonia movil en internet
cursos calidad
cursos de inglés
cursos de verano
Cursos PFP
cursos psicologia
cursos verano
cursos verano tudela
d'Alba
derecho
derecho constitucional español yolanda gomez sanchez
Derecho romano
desarrollo embrionario
descargar compilador de modula-2
Descargar compilador de Modula2
descompresor
dia
diccionario
diccionario de medicina
didáctica del francés
dietetica
diferencia entre autoridad y poder
difusion
diodo
DIPLOMADO EDUCACIÓN SOCIAL
director merida
discreta
diseño de sistemas
DISEÑO Y GESTIÓN
diversificacion curricular
doctorado
doctorado recuperación información
dust in the wind
edipo
educación
educacion social
EL TEATRO EN EL SIGLO XX
electronica digital
empiria
english word TERM
ensamblador
Ernestina
Ernestina de Champourcin
escuela de idiomas de la uned
especialista en ciencia tecnología
estancia hotel
estructura economica españa
estructura tecnologia computadores
estructura y tecnologia

estudios de antropologia
ETB-2/docs.txt/278
ETB-2/docs.txt/295
evaluación prácticas programación
examen
examen programacion II
EXAMENES
examenes
Exámenes de Lógica Matemática
exámenes de lógica matemática
excel
EXPEDIENTE
facultad de psicología-UCM
falgueras
Fatima Gil
fechas de examenes
filologia
filologia inglesa
FISICA INFORMÁTICA
formulario actividades deportivas
formularios
formularios
formularios congresos
formularios Investigacion
formularios mantenimiento
foro del gran hermano
fst 3.0
fst 4.0
fst
fuentes
funciones de tipo carácter
funciones de tipo matemática
galilea
García Cabrero
garcía falgueras
geografia europa
george basalla
góngora
google
gorraiz
grupo de procesamiento del lenguaje
guía
GUIA CURSO DERECHO 2001-2002
guia del curso
guía didáctica de programación ii
guia programacion II
Higinio Mora
historia antigua
historia de la impresora laser
historia del procesamiento del lenguaje natural
horarios escolares

ia
ICAC
identidad masculina
Idiomas
imagen
informacion
Informática
INFORMATICA DE GESTION UNED MADRID
informatica II
ingenieria del software
ingeniería del software
iNTELIGENCIA
interruptor conmutado
INTRODUCCIÓN
investigación ecológica
investigaciones
investigaciones sobre empleo universitario
investigation
inyección
Java
java
Javier San Martín
juegos
juegos
julio gonzalo
Julio hernández rodríguez
junta gobierno 27 octubre 2000
LENGUAJE NATURAL
lerpia
Lexicografía
ley de jurisdicción
libros
Licenciatura en documentación
Linux
lisp
lisp en PLN
lizcano
lógica matemática
LORENZO LUZURIAGA
Luis Fernández Rodríguez
Luis J. Fernández
m2.exe
malaga
manual mathematica
mapas
mapas europa
maps europa
maquina
marta de la cuesta
master en gestion medioambiental y calidad
matematicas

materia de Bretaña
Material
mathcad
matriz funcional
medios de transporte
merelo
metaforas
metal
Metalotecnia
Micro Cap the student edition
miguel mendez
MIGUEL PEÑASCO VELASCO
minguet
MIRA MIRA
modelo de curriculum
modelo plazas profesor encargado
modelo plazas profesor tutor
MODULA 2
modula 2
MODULA
modula
modula programa
Modula-2
modula-2
modula2
modulo cadena
modulos
mohedano
monos
MORILLA
Negociado de Alumnos
niveles de autoridad
noetheriana
notas
notas de programacion I
notas derecho
notas examenes
notas practica
Novalis
obtener modula 2
opsiciones
origen del teatro
Ostenero
paau+extranjero
PALO
PAMPLONA
pamplona
partes del teatro
pascal
Pauling
pdf

pensamiento creativo
peralta
periodismo
pinturas cnossos
PLAN DE ESTUDIOS
Planck
Planes de estudio
plazas profesor
pln
poemas
poesía
Postcript
postgrado unión europea
Postscript
postscript
práctica 2000/2001
PRACTICA
practica
práctica
Practicas
practicas
practicas de programación II de cursos anteriores
practicas programacion
Practicas Programacion
practicas programacion
practicas programacion II
prácticas programación II
precios
preinscripción educación infantil
problemas de lenguaje
profesores
programa practicas programacion II 2001
programacion 2
programacion
Programación DLL
PROGRAMACION I
PROGRAMACION i
programacion I
programacion i
programacion II
programacion ii
programación II
programación ii
programación III
programacion2
programacion_i
programacionII
programacionii
propuesta educativa
Proust
proxy

prueba de grabación
psicologia
Psicología
psicopedagogía
Publicación de la Convocatoria
Quafe 80
quierotv
quimica
ramon peiro
Ramón Sainero
razonamiento analógico aprendizaje
recuperación de información
recuperación de información conceptual
redes
regulación automática
rehabilitacion y seguridad
relevance theory and communication
Reloj de Arena
reloj de sol
representante de profesores-tutores
residuos
respuestas mercantil I
RESULTADOS
resultados+exámenes
riomaior
ROSA PEÑASCO VELASCO
rubio
salinger
salvador Galán
SALVADOR GALÁN RUIZ-POVEDA
santamaria verdejo
Santander Díaz
sap
seguimiento de instrucciones y reglas psicología
selectividad
seminario uned
sense 8
sentido de pertenencia
servidor proxy
sexual dimorphism
signos griegos
Simetría
similitud lenguajes
sistema endocrino
sistemas area de caja banco
sistemas operativos
sociedad esclavista
sociolectos
solicitud ampliación carreras
Soluciones de Fundamentos Físicos de la Informática
soluciones psicología personalidad junio 99

steed
tajetes
teatro isabelino
teleuned
temas sobre redes neuronales
TEORIA DE AUTOMATAS
teoria de automatas
teoría de ejemplares
teorías de aprendizaje
test aptitud verbal
tetris
tiempo
tim read
topografia
topografía
trabajo
Tudela
tutores
tutorias virtuales
uned
uned navarra
UNED PAMPLONA
universidades
uso de la y
vacaciones
Valdo
VERDEJO
Verdejo
veterinaria
vicerrectorado de alumnos
vicerrectorado de metodología
visual basic
wainu
waterpolo
WebCT
webct
winzip

Anexo II: Consultas de sesiones en WTB que empiezan con la exploración de un término y a continuación terminan con la exploración de un documento

CONSULTA ORIGINAL	TÉRMINO EXPLORADO
"Diseño material multimedia"	diseño y elaboración de materiales multimedia
"ligas estaticas"	liga
ALUMNOS ADMITIDOS	alumnos admitidos
analitica de terrenos	terreno
Anselmo Peñas	anselmo lorenzo
Arreglo Darlington	darlington
arte conceptual	arte conceptual
AUTORIZACIONES COMPRA	autorizacion
becas	carta de solicitud de beca
calificaciones_programacion	calificación del ejercicio de programación
caro baroja	caro
carrera de historia	asignaturas de la carrera de geografía e historia de la uned
catedra unesco	cátedra unesco
celina PLN	celina
centro asociado de las rozas	acceder al centro asociado de la rozas
centro asociado rozas	centro asociado de las rozas
centro asociado tres cantos	margen de los relacionados con la sede central y los centros asociados
cine	ciclo de cine uned
compilador modula 2	compiladores de módula
COMPILADOR	compilador de pascal
compiladores +modula 2	compiladores de módula
concepto de simetria molecular	aprehensión de los conceptos de operación y elementos de simetría molecular
conicas	conicas
constructive	constructive
consulta de calificaciones	consulta de calificaciones
convalidacion de estudios	convalidación de estudios
CONVALIDACIÓN	convalidación de estudios
correcion del examen de programacion i	exámen de la segunda semana de programación iii
CREACION DE VALOR	creación de valor
crespo	crespo del arco
CSI & Director	csi

Cuadernillo de Prácticas	cuadernillo de prácticas
cuardenillos de prácticas de programacion	práctica de programación
cuentos para jóvenes	point to a child
cultivo de tejidos vegetales	introducción de cultivos nuevos
cultivos de invernadero	invernadero
Curso 1999	curso 1999
curso de lengua catalana	cursos de idiomas
cursos alimentación	curso de alimentación y salud
cursos de catalan gratis	ejercicio de redacción en catalán
cURSOS DE Idiomas	cursos de idiomas
Definición de mitología	definición de conceptos
del pino artacho	pino artacho
derecho penal	derecho penal ii
deterioro del planeta	desajuste entre recursos y población en el planeta
DICCIONARIO LENGUA ESPAÑOLA	diccionarios de la lengua española e inglesa
dietas	dietas para los profesores
diodos	tipos de diodos
discriminación condicional de segundo orden	discriminaciones condicionales
doctorados	cursos de doctorado
ects pais vasco	ects
El lenguaje de programación javatm	lenguaje de programación
Electrónica Analógica	circuitos electrónicos analógicos
escribano rodenas	juan josé escribano ródenas
Esculturas griegas	escultura y arquitectura
estatuto de la universidad	artículo 35 de los estatutos de la universidad
examenes educacion corregidos	colección exámenes
examenes programacion	examenes
EXAMENES SEPTIEMBRE	éxamenes septiembre
examenes	exámenes resueltos
EXPEDIENTE	expediente
f de snedecor	snedecor
Fania Herrero González	herrero gonzález
fátima gil	gil ferro maria fatima tl
feo	rodríguez feo
FIRST CERTIFICATE	first certificate
FORMULAS LOGICAS	logicas
guia del curso de programacion II	guía didáctica de programación ii
guía del curso	guía del curso
GUIA DIDACTICA PROGRAMCION II	guía didáctica de programación ii
guia didacticacprogramacion II	guía didáctica programación ii
GUTIERREZ MELLADO	universitario general gutiérrez mellado
habilidades motoras basicas	habilidades motoras
habilidades motrices basicas	motriz

HABILITACIONES	cursos de habilitación para profesionales
hardcore	hardcore
home page of programming II	programación ii information page
Huici Urmeneta	vicente huici urmeneta
I02	modelo i02
i07	modelo i07
Ideas principales y secundarias	ideas principales y secundarias
imidazol	evaluación de derivados de imidazol como agentes de constraste para imagen de ph
information retrieval	recuperación de información y lenguaje natural information page
information retrieval	recuperación de información
inteligencia artificial	departamento de inteligencia artificial
juan josé escribano	juan josé escribano ródenas
JUEGO DE PRUEBAS	juego de pruebas
juego simbólico	juego simbólico
julita verdejo	julita
La depuración del profesorado durante el franquismo	depuración del profesorado durante el franquismo
La Sociedad Económica Matritense de Amigos del País	sociedades económicas de amigos del país en la españa del siglo xviii
last course trouble	deranger
learning	language learning
lectura fotografica	lectura óptica
lista_de_correos	lista de correo
lobo	lobo
LSI	departamento de lsi
magisterio infantil	adaptación para diplomados en magisterio
mapas geograficos	mapa topográfico
master en economía	master en unión
matricula gratuita cursos enseñanza abierta	ayuda para cursos matrícula abierta y programa de formación del profesorado
medida	medida
memoria gasto	memoria explicativa y autorización del gasto
metrica	métrica comparada
microeconomia	microeconomia
miguel angel cordova	ángel córdova morales
miguel montano fernandez	miguel fernández
milagros rodriguez	milagros rodríguez olcina
Mitología	mitologia
MODELO INFORMATICO	modelo informático en general
modelos de examenes	modelo para solicitar examenes
modula 2	source installation of modula

modula-2	installation of modula
modula2	modula2
Modula	modula
modula	modula
monitoring	monitoring
movimiento ondulatorio	movimiento ondulatorio
natural language processing	general sobre la problemática y técnicas de la comprensión automática del lenguaje natural
news	news específico de la asignatura
non-Euclidean geometry	returns the geometry
nuevos planes de estudio	implantación de los nuevos planes de estudio
ogando	profesor francisco ogando pasa a ser
Optimización	optimizacion
ora	ora
organos de gobierno	organos de gobierno
Ortiz de la Guía	ortiz de la guia manuel 28800
Ortiz Guía	ortiz de la guia manuel 28800
ozone hole	degradación de la capa de ozono
Pamplona	centro asociado de la uned de pamplona
pamplona	centro asociado de la uned de pamplona
pamplona	uned de pamplona
pas centros asociados	subvenciones a distintos centros asociados por la matrícula del pas de dichos centros
pastanaga	zanahoria
plan nuevo	plan nuevo
plazo de matricula 2001/2002	comienzo del plazo para la presentación de solicitudes de matrícula
Plazo de matrícula abierta	abierto plazo de matrícula por internet curso
pracaticas programacion	programación ii
practicas informatica gestion	informatica de sistemas
preinscripción	plazo de preinscripción
prestamo de libros	incumplimiento de los plazos establecidos para la devolución de los libros en préstamo
primary school	menú principal
problemas	problemas resueltos
procesamiento de lenguaje natural	independent of natural languages
procesamiento online	procesamiento de datos
programacion ii+ septiembre2000	programación ii
programacion II	asignatura de programación ii y de yolanda calero en las labores de programación de las herramientas

programacion II	programación ii
programación	programacion
prueba sin codQ número 2	curso un gran número de preguntas de autoevaluación y pruebas de evaluación
Pruebas de selectividad de matemáticas	guia de matematicas pruebas
ps	postscript
radio lectura imagen	programa de radio con la imagen del profesor
Ramon escuder	escuder cabrejas
rayuela	rayuela
RECTOR	rector
recuperación de information	información not recuperación
reforma social	actuación del instituto de reformas sociales
registro personal	registro de personal
rejuvenecido	rejuvenecer
Renta personal	renta personal
sai	sai
salario profesores	salarios de los funcionarios para el año
segundo ciclo economia	acceso a segundo ciclo de ade y economía
simetria molecular	grupos y la simetría molecular
solicitud autorizacion gasto	memoria explicativa y autorización del gasto
solicitud de comisión de servicios y licencias	recurso contra una multa de tráfico hasta la solicitud de una licencia para realizar obras en una vivienda
tejero	luis tejero
teleprocesos	sistemas especializados en redes de teleproceso
tiempo y sociedad	social y tiempo histórico
transición	crisis económica y transición a la democracia
TRASLADO DE EXPEDIENTE	solicitar traslado de expediente
uned	uned
vamos a ver	dar una visión general sobre la concepción actual del universo
vicente huici	vicente huici urmeneta
visor Postscript	intérprete de postscript y un visor
wainu	wainu
WEBCT	acceso a webct
yoli caro calero	yolanda calero caro