Creación de un recurso textual para el aprendizaje del inglés[1]

Irene Castellón Masalles *
Ana Fernández Montraveta **
Glòria Vázquez García ***

*Departamento de Lingüística General
Universitat de Barcelona
Gran Vía de les Corts Catalanes, 585
Barcelona 08320
Tel: 934034695
Fax: 933189822

**Depatamento de Filología Inglesa y Germanística
Universitat Autònoma de Barcelona
Emprius, 2
Sabadell, 08202
Tel: 937287704
Fax: 937287727

***Depatamento de Filologia Inglesa y Lingüística
Universistat de Lleida
Pl. Víctor Siurana, 1
Lleida, 25003
Tel.: 973703140
Fax: 973702170

Resumen

En este artículo presentamos el corpus CPG (Corpus Paral·lel GRIAL[2]), que se ha creado como parte de un proyecto de innovación docente con el objetivo de mejorar los procesos de enseñanza/aprendizaje usados en la asignatura de Inglés Técnico, que se cursa en el primer curso de Ingeniería Informática. Dicho recurso es un un corpus paralelo formado por 1.031.911 palabras para el inglés, 393.684, para el catalán, y 831.903, para el español, y pertenece al registro técnico, más específicamente, al dominio de la informática. Todos los textos que configuran el corpus se han anotado a nivel morfosintáctico, lo cual permite realizar consultas más complejas que solamente el contexto léxico de una palabra.

Palabras clave: corpus paralelos, etiquetaje morfosintáctico, recursos docentes.

Abstract

In this paper we present the CPG (Corpus Paral·lel GRIAL) which has been built as part of a teaching innovation project aimed at enhancing the teaching and learning processes in an introductory Computer Science course.

[1] Este proyecto se ha realizado gracias a la ayuda de la Generalitat de Catalunya (194 MQD 2002).
[2] Grup de Recerca Interuniversitari en Aplicacions Lingüístiques.

This resource is a parallel corpus made up of 1.031.911 English words, 393.684 Catalan Words, and 831.903 Spanish words from a technical register, more precisely, from the computer science domain. All texts have been morpho-syntactically tagged, allowing for high level linguistic queries.

Keywords: parallel corpora, morpho-syntactic tagging,

Introducción

El proyecto ha consistido en la creación de un corpus paralelo, es decir, una colección de textos disponibles en diferentes lenguas, en este caso, inglés-catalán-español. Dicho corpus está formado en total por 2.257.498 palabras y pertenece al dominio de la informática. El objetivo último es el uso de esta herramienta como recurso didáctico para la impartición de la materia de inglés Técnico a los estudiantes de primer curso de Ingeniería Informática.

La característica principal de un corpus paralelo es que contiene textos en diferentes lenguas que son traducciones de la misma fuente. Ello permite establecer comparaciones lingüísticas muy interesantes que pueden explotarse en diferentes campos. En nuestro caso, como dicho recurso se utiliza en el ámbito docente, la ventaja principal de la paralelización se basa en que los alumnos pueden comparar los usos gramaticales de su lengua materna (ya sea catalán o castellano) con los de la lengua de aprendizaje, el inglés. Por otro lado, el uso de corpus como fuente de información hace que los alumnos trabajen con el uso real de la lengua y no con ejercicios y vocabulario prefabricado que quizás no les serán de utilidad con vista a su futuro profesional (Fernández y Coll 2000).

Además, como el corpus ha sido anotado a nivel morfológico, la explotación de dicho recurso es de especial interés para el profesor a la hora de preparar material didáctico.

La metodología utilizada se fundamenta en la denominada 'task-based-learning' (aprendizaje basado en tareas – J. Willis 1996), según la cual la lengua inglesa deja de ser una finalidad en si misma para convertirse en una herramienta para llevar a cabo la tarea marcada por el profesor.

En este artículo se describe el proceso de elaboración del recurso y algunas cuestiones preliminares relacionadas con su explotación. Se ha estructurado del siguiente modo: en el apartado 2 se presenta el proceso de constitución del corpus; en el siguiente apartado (3) se presenta la interfaz de explotación y sus posibilidades; posteriormente, dedicamos el apartado 4 al uso docente del recurso y, finalmente, en el apartado 5, se exponen las líneas futuras de continuación del proyecto.

2. Constitución del corpus

La constitución del corpus ha sido la primera tarea del proyecto. El objetivo de esta tarea ha sido doble: por un lado, nos interesaba un corpus con diversidad textual y lingüística; por otro, y pensando en su explotación, el corpus requería de anotación morfológica para poder realizar búsquedas más refinadas sobre los textos. El proceso de constitución del corpus se llevado a cabo en tres fases: a) la recopilación de los textos, b) la paralelización y c) la anotación.

2.1 Recopilación de los textos

Las fuentes que configuran el corpus son de diferentes tipos según su formato original. Por un lado, hemos utilizado fuentes electrónicas, fundamentalmente documentos extraídos de la web y algunos artículos de revistas del área, y, por otro lado, hemos utilizado fuentes textuales en papel, escaneando los textos para convertirlos al formato digital.

Por lo que se refiere al tipo de fuentes según su contenido, hemos recopilado diversos tipos de textos para disponer de un corpus que recoja las peculiaridades del discurso técnico de cada uno de los diferentes registros dentro del dominio de la informática, con el objetivo de que el corpus fuera lo más variado posible en cuanto a estructuras. Así, se han empleado tres tipos de textos: manuales de software, artículos de revistas de informática y libros. En la figura 1, presentamos la proporción de los textos ingleses según el tipo de documento.

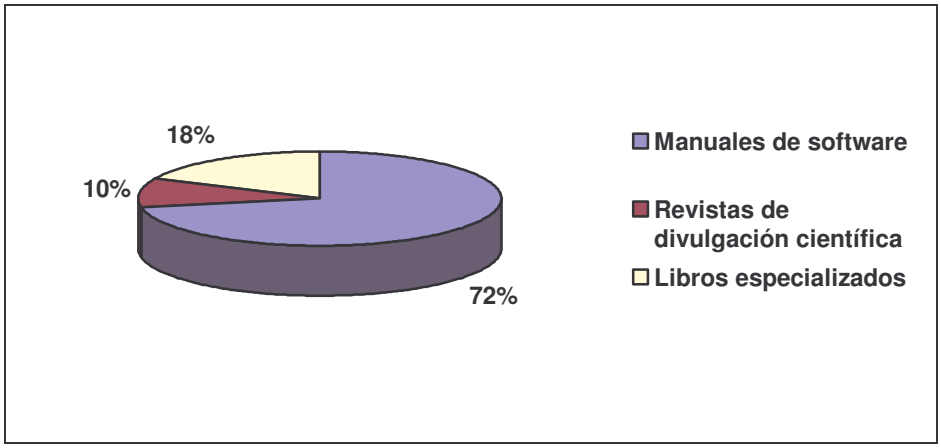

Figura 1. Distribución según el tipo de texto para el inglés

Como se puede observar, la mayor parte de la documentación se compone de manuales, que es el tipo de documento más abundante actualmente en este ámbito y, además, se obtienen libremente. Así mismo, es el tipo de documento que más fácilmente hemos obtenido para el catalán.

En la tabla 1 se indica la distribución de lenguas. El motivo de que ésta sea irregular es que la disponibilidad de textos bilingües es diferente en cada lengua. Pese a esto, este proyecto es abierto y la arquitectura del sistema de gestión de textos está pensada de forma que se pueda enriquecer la base de datos y equilibrar el volumen de palabras de

las diferentes lenguas. Respecto a los textos procedentes de revistas, el total (10 % del corpus) corresponde a las monografías cedidas por Novática[3].

Lengua	Total de palabras	Porcentaje
Inglés	1.031.911	45,7%
Español	831.903	36,8%
Catalán	393.684	17,4%

Tabla 1. Distribución según lenguas

2.2 Paralelización del corpus

La paralelización del corpus consiste en asociar las traducciones de los documentos de las diferentes lenguas alineando los fragmentos equivalentes entre éstas. Los fragmentos son, en general, oraciones (cadenas de palabras entre puntos), o bien unidades más pequeñas (sintagmas), en el caso de que no existan oraciones, por ejemplo, títulos o listas de nombres propios.

En el caso de los textos trilingües, la alineación se ha llevado a cabo entre el inglés y el español, por un lado, y entre el inglés y el catalán, por otro, ya que la herramienta utilizada sólo permite la paralelización entre dos lenguas. El resultado final es el mismo (la alineación de los tres textos) aunque la elaboración es más costosa.

La fase de la paralelización se ha realizado de forma totalmente manual para asegurar la calidad del recurso utilizando el programa "Transuite Align 2000". En la figura 2 podemos ver el aspecto de la interfaz con un fragmento de texto alineado entre inglés y

[3] Queremos agradecer a Novática la cesión de dichos textos.

español. El resultado de la alineación es una base de datos donde apararecen los dos textos en diferentes campos y las oraciones o fragmentos ocupan los registros.

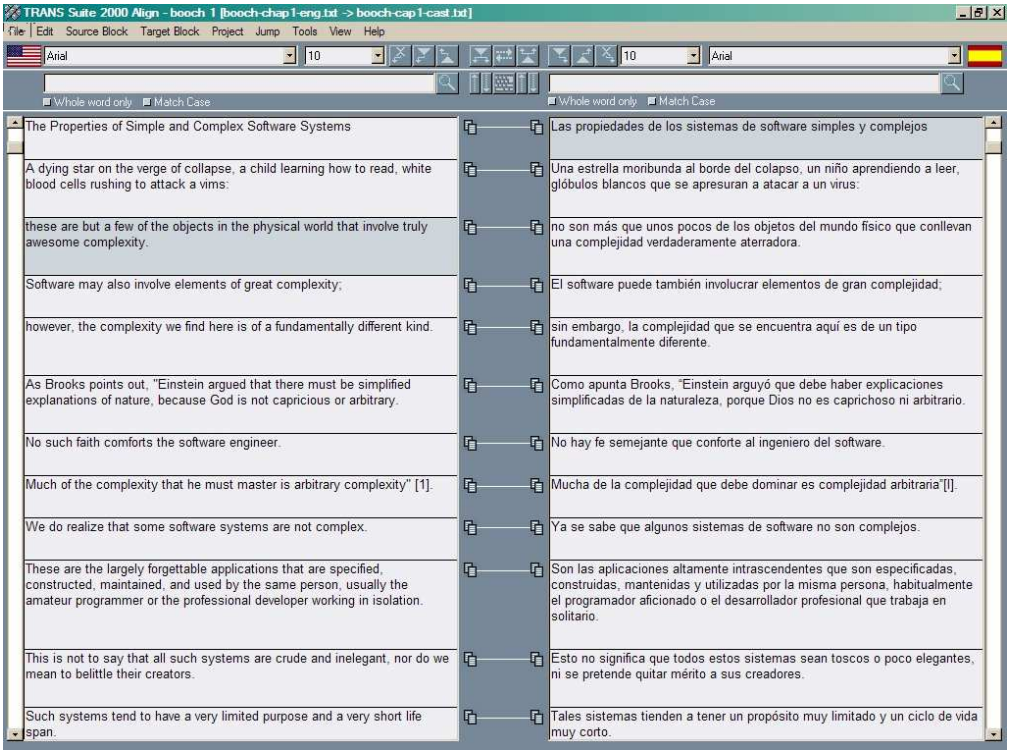

Figura 2. Interfaz de paralelización de "Transuite Align 2000"

En algunos casos encontramos traducciones totalmente literales del documento fuente[4]. Sin embargo, estos casos son los menos frecuentes, ya que es común que determinados fragmentos del documento original no tengan una correspondencia con el documento destino, o a la inversa; en estos casos hemos optado por alinear dicho fragmento con un operador nulo que indica la no existencia de correspondencia, para poder conservar el documento original.

También ocurre que, aunque dos textos tengan el mismo número de fragmentos, éstos se pueden presentar en diferente orden en cada uno de los textos, puesto que la disposición

[4] En este corpus sólo hay traducciones literales en los textos provinientes de Softcatala, ya que en su gran mayoría son procedentes de manuales y se componen de oraciones imperativas (órdenes) muy consolidadas en el área informática.

de las palabras y los sintagmas más natural varía de una lengua a otra. Ello ha motivado algunas alineaciones entrecruzadas que modifican la ordenación de los fragmentos en la lengua destino. Un ejemplo claro de este tipo de alineación lo constituyen los textos de Novática, que son monografías de autor, donde la traducción no sigue exactamente la estructura del texto original.

Anotación del corpus

El siguiente proceso que hemos aplicado a los corpus, una vez paralelizados, es la anotación a nivel morfológico. Esta anotación se ha realizado de forma automática y ha sido necesario realizarla de forma independiente para cada lengua. Aunque idealmente hubiera sido conveniente utilizar el mismo tipo de analizador para la etiquetación de los textos en las tres lenguas, sólo ha sido posible usar la misma herramienta para el español y el inglés. El sistema de análisis utilizado para estas lenguas ha sido "Connexor" (www.connexor.com) y para el catalán se ha utilizado "Free-Ling" (http://www.lsi.upc.es/~nlp/freeling/).

Uno de los inconvenientes es que cada uno de estos analizadores utiliza un conjunto de etiquetas distinto, por lo que, en un estadio posterior, se han establecido las correspondencias necesarias entre estos conjuntos. Por lo que se refiere al nivel morfológico, uno de los motivos de estas diferencias es que las tres lenguas difieren en algunos aspectos; por ejemplo, en inglés, los nombres, adjetivos y determinantes no tienen género y en español y catalán sí. Además, la misma herramienta codifica información diferente en función de la lengua aunque no hay un motivo estructural; por ejemplo, "Connexor" codifica la información referente al caso para el inglés y no para el español, como se puede observar a continuación:

Engineer: N NOM SG Ingeniero: N MSC SG

No obstante, las diferencias más notables por lo que se refiere al conjunto de etiquetas se da en relación al catalán, ya que en "Free-Ling" la nomenclatura es totalmente diferente. Así, para etiquetar un nombre masculino singular como "enginyer" se utiliza la nomenclatura NCMS000.

Todo ello nos ha llevado a reducir el nivel de detalle de estas categorías. Así, la correspondencia se ha establecido en base a las categorías genéricas (verbo, nombre, etc.) y se ha prescindido de las subcategorías que aportan la información relativa al número y al género, por ejemplo. Esta generalización no tiene repercursiones importantes en la consulta de los datos, ya que este tipo de información no es estrictamente necesario para hacer las búsquedas que se han previsto.

Por otro lado, los resultados de ambos analizadores son diferentes en función de la información sintáctica, que "Connexor" proporciona y "Free-Ling" no. Así, los textos ingleses y españoles contienen información sobre las dependencias que se establecen entre las diferentes unidades de la oración. Esta información puede ser muy útil y no descartamos utilizarla más adelante, sin embargo, convendría evaluar la bondad del resultado y, en consecuencia, en la versión actual de la interfaz no se contempla.

Formalmente, la información que se visualiza en la anotación de los textos en español e inglés no es uniforme con respecto a la que se visualiza en los textos en catalán ni se expresa de la misma forma. Así, "Conexor" utiliza el formato XML y proporciona tanto información mofológica, concretamente, la información sobre el lema (<lemma>) y la categoría (<Morpho>), como sintáctica: la dependencia (<depend>), asociada a los elementos no nucleares, y la función sintáctica (<syntax>).

En la figura 3, podemos observar cómo se representa la frase del inglés "*Logically, the threat level of a virus can vary"*. Como se puede observar, la palabra *level* tiene como lema "*level*", como categoría "N NOM SG" nombre nominativo singular), depende de

la unidad 416 (*can*), es decir, del verbo[5], y tiene asociada como función la de sujeto (@SUBJ %NH) de dicho verbo.

```
<sentence id="29">
<token id="w408">        <text>Logically</text>    <lemma>logically</lemma>
          <tags><syntax>@ADVL %EH</syntax> <morpho>ADV</morpho></tags></token>
<token id="w409">        <text>,</text>    <lemma>,</lemma></token>
<token id="w410">        <text>the</text> <lemma>the</lemma>     <depend head="w412">
det:</depend>    <tags><syntax>@DN&amp;gt; %&amp;gt;N</syntax>
<morpho>DET</morpho></tags></token>
<token id="w411">        <text>threat</text>        <lemma>threat</lemma> <depend
head="w412">attr:</depend>       <tags><syntax>@A&amp;gt; %&amp;gt;N</syntax> <morpho>N NOM
SG</morpho></tags></token>
<token id="w412">        <text>level</text>        <lemma>level</lemma>   <depend
head="w416">subj:</depend>       <tags><syntax>@SUBJ %NH</syntax> <morpho>N NOM
SG</morpho></tags></token>
<token id="w413">        <text>of</text>   <lemma>of</lemma>       <depend
head="w412">mod:</depend>       <tags><syntax>@&amp;lt;NOM-OF %N&amp;lt;</syntax>
<morpho>PREP</morpho></tags></token>
<token id="w414">        <text>a</text>    <lemma>a</lemma>        <depend
head="w415">det:</depend>       <tags><syntax>@DN&amp;gt; %&amp;gt;N</syntax> <morpho>DET
SG</morpho></tags></token>
<token id="w415">        <text>virus</text>        <lemma>virus</lemma>   <depend
head="w413">pcomp:</depend>   <tags><syntax>@&amp;lt;P %NH</syntax> <morpho>N NOM
SG</morpho></tags></token>
<token id="w416">        <text>can</text><lemma>can</lemma>     <depend head="w417">v-
ch:</depend>     <tags><syntax>@+FAUXV %AUX</syntax> <morpho>V
AUXMOD</morpho></tags></token>
<token id="w417">        <text>vary</text>        <lemma>vary</lemma>
```

Figura 3: Análisis morfológico de "Connexor" para el inglés.

El analizador morfológico para al catalán (Carreras et al. 2004) presenta la estructura de la anotación mediante tripletas formadas por la forma, el lema y la categoría. Para expresar esta última se utiliza un conjunto de etiquetas estandarizadas, propuestas por el grupo EAGLES (1996). En la figura 4, se muestra un análisis de la frase en catalán "*Cada paquet de la distribució té executables"*. Como ejemplo, podemos observar que en el análisis para la forma *la* se indica el lema correspondiente ("el") y su categoría ("DA0FS0": determinante femenino singular).

[5] La anotación utilizada en "Conexor" se basa en el análisis de dependencias, donde el verbo es el núcleo de la oración y del cual dependen los demás constituyentes básicos de la oración.

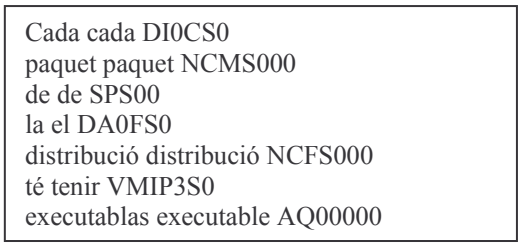

Figura 4. Análisis morfológico de "Free-Ling"

Una vez anotados todos los textos, se reconstruyó mediante un proceso automático la paralelización realizada con los textos en la etapa anterior, para poder disponer de la alineación con anotación.

3. Construcción de la interfaz[6]

La interfaz diseñada permite realizar búsquedas en el corpus tanto sobre el texto como sobre la información añadida en el proceso de anotación. Así, por un lado no sólo se pueden realizar búsquedas de frases donde aparezca la palabra exacta (p.e. *computerize*), sino que también se pueden buscar todas las formas asociadas a un lema (*computerize, computerizes, computerized*). Este tipo de búsqueda presenta importantes ventajas sobre todo para el catalán y el español, lenguas en las que la flexión es más rica.

En la figura 5 se puede apreciar el aspecto de la interfaz de consulta. El tipo de anotación que contiene el corpus permite refinar la búsqueda en casos de ambigüedad, permitiendo asociar a una forma o a un lema su categoría para discrimar casos que no son parte del objetivo de búsqueda (por ejemplo, *place* puede ser verbo o nombre).

Por otro lado, se puede ampliar la búsqueda a más de un elemento, posibilitando las búsquedas a nivel sintáctico. Por ejemplo, se puede interrogar al sistema sobre patrones gramaticales a través de la combinación de más de una categoría (verbo + nombre) o de una categoría y un lema o una forma (verbo + *bit/bits*).

[6]http://grial.uab.es:8080/exist/cpg/

En el ejemplo de la figura 5, se interroga al sistema sobre la existencia de *computer* como sustantivo seguido de otro sustantivo. Por último, el sistema permite seleccionar los textos sobre los que se realizará la búsqueda, lo cual posibilita prescindir en determinadas ocasiones de algún tipo de textos (como por ejemplo los manuales) que tienen estructuras muy repetitivas y que para determinadas búsquedas no pueden aportar resultados.

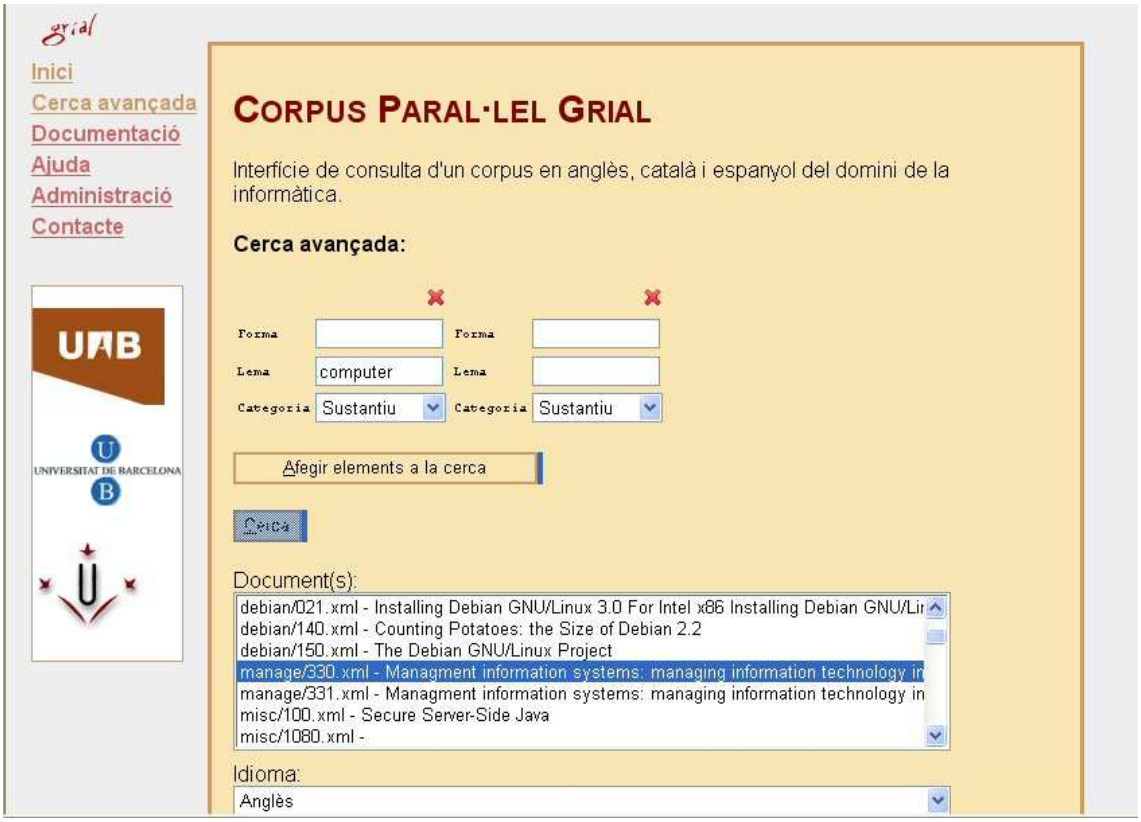

Figura 5: Aspecto que ofrece la interfaz de consulta de corpus

El sistema presenta en los resultados siempre las tres lenguas (excepto en el caso de que no haya fragmento equivalente). Podemos ver un ejemplo del resultado obtenido de la búsqueda antes especificada en la figura 6.

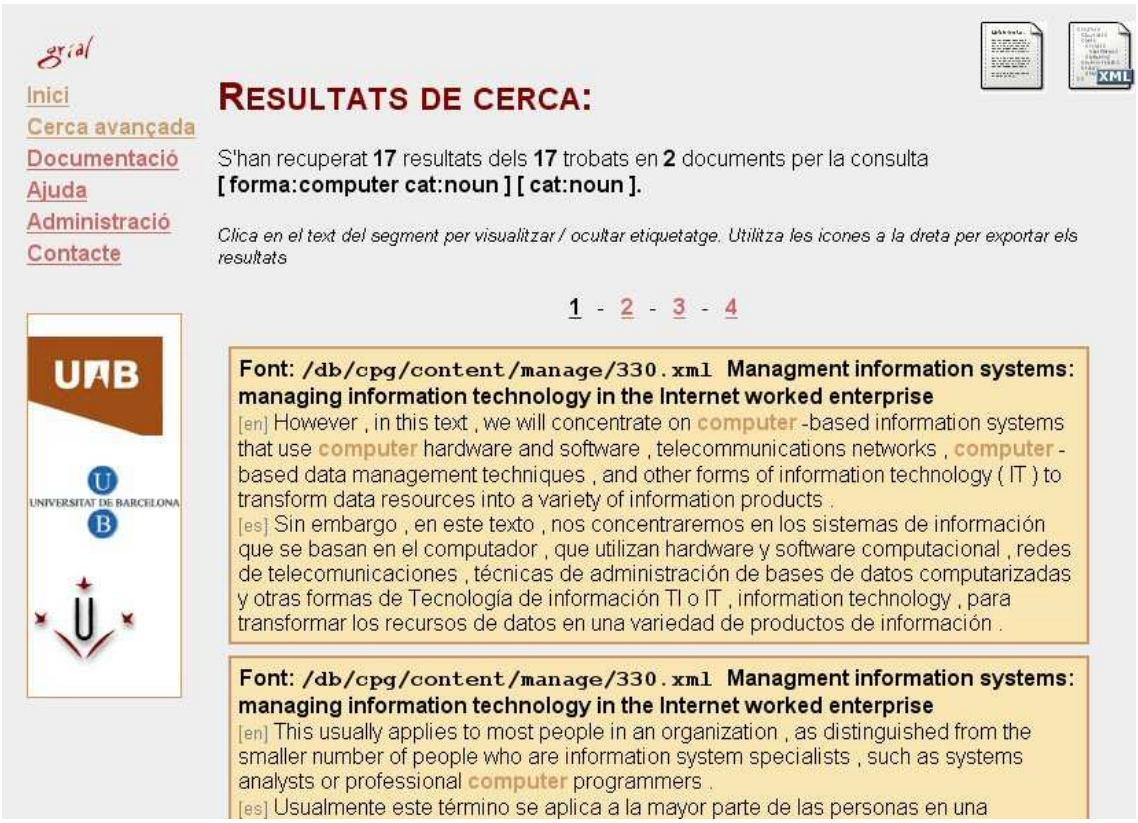

Figura 6. Resultado de una búsqueda

4. Creación de materiales y diseño de actividades

El recurso creado presenta importantes ventajas para su explotación, tanto por las características del corpus, que está anotado a nivel morfológico y contiene también algún tipo de información sintáctica, como por las de la interfaz de consulta, que permite realizar búsquedas de más de un elemento

Así, en primer lugar, es una herramienta muy útil para el profesor para extraer material para la creación de ejercicios, preparación de textos de lectura, ejemplos, etc. En este sentido el profesor dispone de una fuente altamente valiosa que le permite preparar tareas para ilustrar los aspectos gramaticales más importantes y trabajar con determinadas funciones del lenguaje según el tipo de alumnado.

Como fuente de información permite al docente estudiar peculiaridades del inglés técnico y seleccionar los temas a presentar en clase. El diseño del material y, por tanto, el diseño del curso, está condicionado por el resultado de las búsquedas, puesto que lo que se pretende es acercar el uso del inglés técnico real al alumno y trabajar aquellos aspectos lingüísticos que le son pertinentes.

Por ejemplo, se ha llevado a cabo un estudio del uso de condicionales en los corpus disponibles y se ha observado que, de los tres usos del condicional explicados tradicionalmente en clases de inglés, sólo se considera relevante para el inglés técnico uno de ellos, el denominado condicional de hipótesis reales. Así, el ejercicio práctico para trabajar el condicional se ha centrado en este tipo. Otro caso similar es el de los usos de los verbos denominados modales, ya que las funciones que tienen estas palabras en el registro del lenguaje técnico son más restringidas que los significados que pueden expresar en un dominio general del lenguaje.

A nivel más general, entre los tipos de explotación del corpus para trabajar la gramática del inglés nos gustaría destacar los siguientes:

- Inducción de reglas gramaticales
- Contrastación de la teoría con el lenguaje real
- Estudio de colocaciones léxicas.

Los dos primeros tipos están muy relacionados. En cuanto a la inducción de reglas gramaticales, este tipo de enfoque implica que, en vez de explicar una regla gramatical al alumno, éste sea capaz de formularla por sí mismo a partir de la observación del uso de una palabra en oraciones. Por ejemplo, si el alumno estudia el comportamiento del verbo *want* observará que siempre aparece en tres tipos de construcciones:

- Verbo + sintagma nominal

- Verbo + verbo en infinitivo
- Verbo + objeto + verbo en infinitivo

Otra posibilidad es que el alumno utilice esta información inferida para contrastarla con su propio conocimiento sobre el uso real de este verbo.

Finalmente, por estudio de colocaciones léxicas nos referimos a las preferencias de coaparición de determinadas palabras, por ejemplo, qué preposiciones aparecen con qué verbos o bien determinados compuestos nominales frecuentes en el discurso técnico.

En lo referente al análisis textual se trabajan aspectos como los siguientes:

- Estructura del texto, del párrafo y del discurso técnico
- Funciones del lenguaje.

Al recopilar textos completos es posible estudiar, por ejemplo, la estructura de un artículo de revista o de un manual de uso. Por otro lado, también se puede trabajar la organización de la información en el texto, ya que el párrafo suele coincidir con la expresión de una idea principal y detalles referentes a la misma.

En cuanto a las funciones, se trabajan los diferentes registros de uso correspondientes a los diferentes textos, por ejemplo, cómo se utilizan determinados tiempos verbales o construcciones en cada uno de ellos.

A partir de todo lo expuesto, las ventajas que implica para los estudiantes el uso del corpus como recurso dinámico en las clases son evidentes. El hecho de que los materiales se creen específicamente para la tipología de alumnos de Informática ha de permitir que los estudiantes se sientan más cómodos a lo largo del proceso de enseñanza/aprendizaje y que, por tanto, aumente el nivel de motivación y, en consecuencia, dicho proceso tenga más éxito.

5. Líneas de futuro

Aunque oficialmente ha finalizado ya el proyecto, nuestra intención es continuar trabajando dentro de la línea iniciada. En este sentido, y en lo referente al corpus, el proyecto está abierto a la inclusión de nuevos documentos. Ya en el diseño del sistema de explotación se optó por un sistema abierto que permitiera ir ampliando la selección de textos. La ampliación del corpus permitirá equilibrar el recurso en diversos sentidos. Esta ampliación pasa forzosamente por aumentar el número de textos del tipo revista o libro para equilibrar la tipología de textos en el ámbito informático. Además, se pretende igualar, tanto como sea posible, los textos del español y del catalán, en relación con el inglés. También es un objetivo a más largo plazo ampliar el corpus a otros dominios temáticos como por ejemplo el de economía, derecho, etc. y así crear un corpus de referencia del inglés para usos específicos.

Referencias

Carreras, X., Chao, I. Padró, Ll, Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

EAGLES: Evaluation of Natural Language Processing Systems, Final Report: http://www.issco.unige.ch/projects/ewg96/ewg96.html

Fernández, A.; M. Coll (2000). "Integration of computers in the classroom: the use of UNIX". Encuentros, 11, pp. 87-95.

Willis, J. (1996). A Framework for task-based learning. Ed: Longman.