El Corpus Científico-Técnico de la Real Academia Española

ALMUDENA BALLESTER CARRILLO
Real Academia Española

Introducción

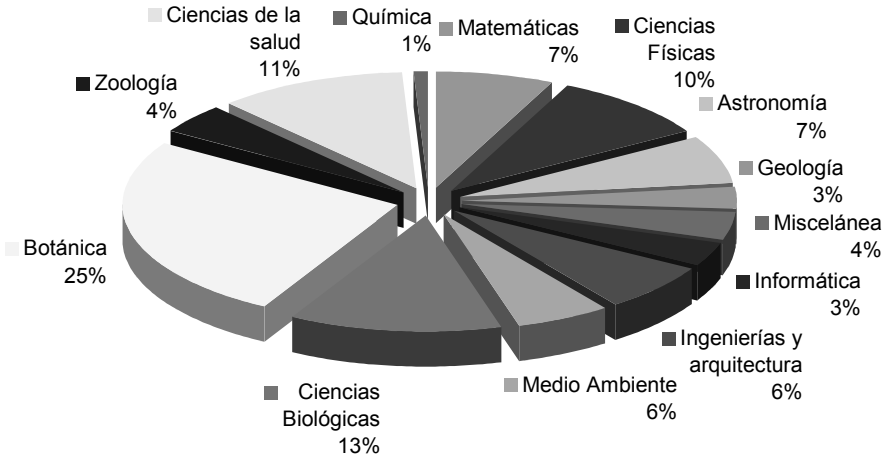
La finalidad del proyecto del Corpus Científico-Técnico (*CCT*) es la construcción y explotación de un recurso textual especializado que permita, por un lado, el incremento y la mejora de los diccionarios y vocabularios de los dominios de la Ciencia y la Tecnología y, por otro, el desarrollo de herramientas lingüísticas orientadas al estudio de diferentes aspectos morfológicos, sintácticos y semánticos del lenguaje científico.

Un corpus especializado como el *CCT* permitirá constatar el uso de términos conocidos, verificar su índice de dispersión y propagación a otros dominios y probar técnicas de reconocimiento automático de nueva terminología. Los resultados que se puedan extraer de su explotación siempre resultarán más fiables que los obtenidos de un corpus *generalista*, que no posee una representación adecuada en sus correspondientes parcelas del conocimiento científico y técnico.

Composición

El corpus en su estado actual lo conforman 1.134 textos, que suman un total de más de seis millones de palabras, agrupadas por dominios. Aunque se pretende un equilibrio temático, se ha optado en esta primera fase por incorporar un gran volumen de textos, producto de acuerdos con diversas instituciones de prestigio, haciendo prevalecer el criterio de «garantía». Por lo cual, en la actual composición destaca sobre el resto de las áreas cubiertas la de Botánica, fruto del acuerdo de la Real Academia de Ciencias con el Real Jardín Botánico - CSIC y la cesión de su extensa monografía *Flora Iberica.*

Asimismo, otras fuentes destacadas son la revista *Investigación y Ciencia* (cuyos números desde 1996 hasta 2003 están integrados en el corpus), los anuarios de la Real Academia de Ciencias y el tratado *Principios integrales de la Zoología*, de C. P. HICKMAN *et al.*



Composición del corpus

Un subcorpus periférico, de aproximadamente 3 millones de palabras extraídas de sitios de internet con perfil técnico y científico, complementa el corpus nuclear en su dimensión de banco de extracción terminológica. Esta «segunda capa» se ve incrementada con artículos u otras publicaciones de revistas científicas, universidades y empresas dedicadas básicamente a la investigación. Entre sus principales componentes, hay que destacar los manuales médicos del grupo MSD, entre ellos, el Manual MERCK completo.

Una de las características principales de todos los textos integrantes de este corpus es que han sido obtenidos directamente de su versión electrónica, con el considerable ahorro de tiempo de verificación y preproceso.

Codificación del *CCT*

Fieles a los estándares europeos de codificación de corpora, el *CCT* ha adoptado la norma XCES (XML) y se adaptó para tal fin una DTD propia.

De los datos contenidos en la cabecera se extrae valiosa información tanto bibliográfica y temática como puramente lingüística o de proceso: número de palabras, información sobre la marcación XML, etc.

El proceso de codificación ha sido realizado semiautomáticamente, valiéndose de las estrategias y programas desarrollados por el Departamento de Lingüística Computacional (DLC) a tal fin. Estas herramientas de transformación parten de los textos en diversos formatos y los vuelcan al formato XCES, con su respectiva cabecera.

El *CCT* también existe como único documento xCesCorpus, para posibilitar su validación y procesamiento conjunto, por medio de una aplicación XML específica.

Control de calidad e intervención manual

La intervención manual es mínima en la mayoría de los casos, y se limita a comprobar y validar la estructura XCES de los textos, así como a corregir pequeños problemas de adaptación de fórmulas científicas y, en general, particularidades de los textos científicos y técnicos.

Procesamiento lingüístico

Se aplican las mismas fases de proceso que para el resto de los corpus que componen el Banco de Datos de la Real Academia Española: segmentación por unidades de análisis, anotación morfosintáctica y desambiguación automática, con las mismas herramientas que ha desarrollado el DLC, si bien, adaptadas a las particularidades de estos textos: reconocimiento de fórmulas, léxico específico, reglas automáticas adaptadas a la estructura sintáctica habitual del lenguaje de la ciencia, etc.

También se dispone de un subconjunto de textos, 50.000 palabras en total, que han sido desambiguados manualmente. Su propósito es doble: comprobar y mejorar la estrategia de desambiguación automática para los textos científicos y constituirse en «prototipo» destinado a verificar las herramientas de extracción de terminología.

Herramientas propias

Scripts de transformación:
Una serie de programas escritos en lenguaje Perl toman como entrada los ficheros convertidos por el procesador Word (si es el caso) a HTML y eliminando la codificación redundante e innecesaria, normalizan la salida. Un primer script elimina comentarios vacíos y algunos elementos de formato y cambia algunas entidades. Un segundo compacta la estructura, incorpora una hoja de estilo CSS y devuelve un documento XML bien formado. Un tercero reasigna la DTD, elimina elementos vacíos y cambia

algunos otros nombres de elementos. Por último, un script de salida limpia, convierte y elimina varios de los elementos que habrán de ser validados. A continuación, el documento, ya con estructura XML, queda listo para ser editado y validado de acuerdo con la DTD de XCES.

Scripts de extracción de información:
Un segundo tipo de scripts vuelca la información contenida en los textos hacia otras estructuras: son herramientas que procesan los glosarios en formato electrónico y crean estructuras XML para su posterior proceso, así como extractores de información de campos de cabecera para crear la nómina textual y extractores de estructuras sintagmáticas y *n-gramas* del corpus.

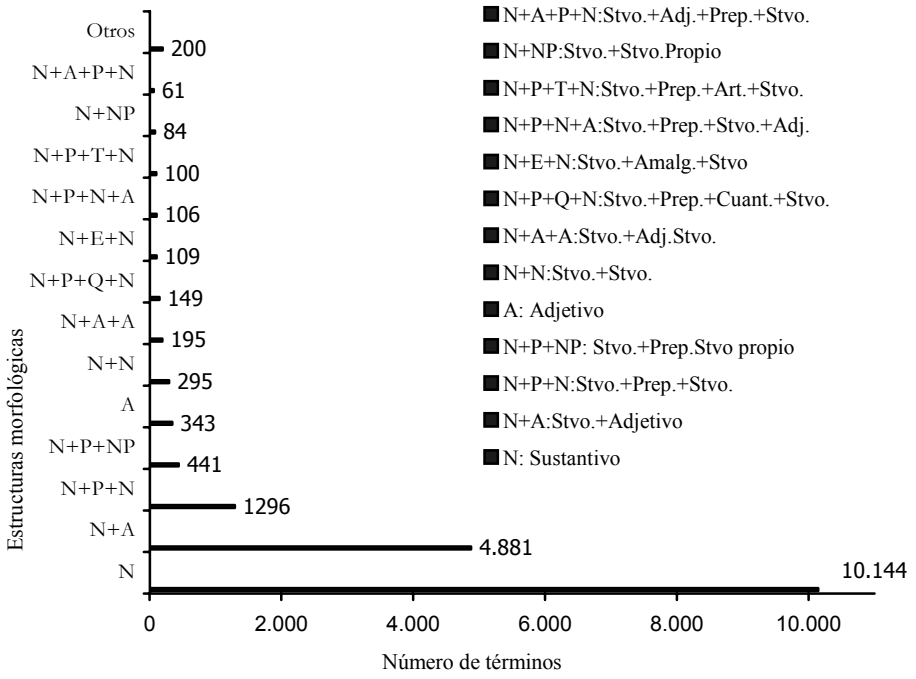
Scripts de visualización:
Son aquellos con los que se crean los documentos visualizables en cualquier navegador estándar; básicamente, contienen información sobre el proyecto y sus documentos. Se usaron conversores de estilo y código HTML, principalmente.

Recursos terminológicos

Como recurso más importante, el *CCT* se vale del *Diccionario Esencial de las Ciencias* de la Real Academia de Ciencias.

En primer lugar este diccionario especializado fue transformado desde su formato original al formato estándar XML. Se desarrolló asimismo una DTD específica y más tarde se procesó el conjunto por dominios, dotando a cada entrada de su morfosintaxis requerida. El resultado supone un estupendo fondo lexicográfico que ha sido una estrategia básica para el reconocimiento garantizado de verdaderos términos en este corpus Científico-Técnico.

A continuación, se muestra la estructura morfosintáctica de los términos contenidos en el *Diccionario Esencial de las Ciencias* (edición de 1999) así como el número de entradas terminológicas que la comparten (de un total de 18.000 procesadas):



Además de esta fuente, del propio corpus se ha extraído un minilexicón de lemas nuevos, que se ve incrementado obviamente a medida que se introducen nuevos textos, y que, a su vez, realimenta el recurso léxico principal.

Se han preprocesado otros glosarios terminológicos, entre ellos:

- el *Diccionario médico etimológico* de MONTEFIORE
- el glosario de Saludalia
- la Neoloteca del Centre de terminología TERMCAT
- el *Diccionario terminológico y de siglas del subsector de lubricantes* de la Asociación Española de Lubricantes
- el *Glosario médico multilingüe* de la DG III de la Unión Europea
- el *Glosario oftalmológico* de Oftalmored

Reconocimiento automático de términos

Un subcorpus segmentado y equilibrado de dos millones de palabras, extraído del propio *CCT* , sirvió de campo de pruebas para un sistema de reconocimiento automático de términos, basado en una precisa combinación de técnicas estadísticas con la extracción de *n-gramas* del propio corpus.

La pretensión de dicho sistema fue no tanto extraer verdadera terminología del corpus, sino *candidatos* de término, que serán más tarde validados por terminólogos y lexicólogos del Instituto de Lexicografía de la Real Academia. Incrementar los diccionarios especializados de esta institución es el objetivo de estos procedimientos.

Basado en tres dimensiones relevantes y no correlacionadas entre sí, los *n-gramas* fueron caracterizados de acuerdo a su:

Cohesión léxica: Una medida de *información mutua* propuesta por YAMAMOTO y CHURCH, GMI (Generalized Mutual Information), que compara la frecuencia de un *n-grama* con la frecuencia esperada si las partes fueran combinadas de manera independiente (para *n-gramas* >1).

Pertenencia a dominio: Medido según IDF (Inverse Document Frequency), —SPÄRK JONES— que calcula la cantidad de información que se da al sistema si se le dice que un documento en cuestión contiene dicho *n-grama*.

Uso especializado: Computado con R (Relative Frequency Ratio), que divide la frecuencia de un posible término en un corpus generalista de referencia, en este caso, un subcorpus literario del *Corpus de Referencia del Español Actual* (*CREA*) de 30 millones, contra la frecuencia en el corpus especializado, *CCT*.

El sistema fue previamente entrenado con un conjunto de términos positivos y otro de no-términos. Se corrigieron los efectos de desvío y de ruido y se introdujeron algunos filtros ortotipográficos y morfológicos para reducir la tasa de error. Finalmente, se aplicó el algoritmo de computación *C4.5* (QUINLAN 1993), que produjo un conjunto de *candidatos* de término.

Muestra:

Término propuesto	Cohesión léxica (GMI)	Pertenencia a dominio (IDF)	Uso especializado (R)	Pertenece al DEC
atracción gravitacional	12.46	5.33	0.009	NO
columna vertebral	14.65	3.74	0.220	SÍ
emisión de positrones	10.70	4.59	0.006	NO
proteína quinasa	11.18	6.91	0.003	SÍ
leucemia mieloide	16.84	4.91	0.013	NO

Secuencias descartadas	Cohesión léxica (GMI)	Pertenencia a dominio (IDF)	Uso especializado (R)	Pertenece al DEC
molécula de agua	5.64	5.33	0.007	NO
multiplicación de las especies	5.00	5.91	0.016	NO
sobre la tierra	2.98	4.91	2.084	NO
de los huesos	- 0.35	4.90	1.018	NO

En los resultados se puede apreciar la inmediata aplicación de estas técnicas de reconocimiento de nuevos términos para el incremento de la lexicografía especializada.

Perspectivas futuras

Será necesario, en primer lugar, el incremento sistemático del número de textos del corpus, para lograr una representación *fiable* de cada parcela científica o técnica abarcada. Conjuntamente a ello, se dotará al corpus del equilibrio temático oportuno.

Se pretende, asimismo, seguir desarrollando técnicas de extracción automática de términos, así como herramientas que hagan más fácil el acceso y explotación por parte de los usuarios finales de este corpus.

Bibliografía

QUINLAN, J. R. (1993) *C4.5: Programs for Machine Learning,* Morgan Kauffmann, San Mateo, CA

SPÄRK JONES, Karen (1973) «A Statistical Interpretation of Term Specificity and its Application in Retrieval», 11-21 en *Journal of Documentation* 28 (1)

YAMAMOTO, Mikio / Kenneth W. CHURCH (2001) «Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus», 1-30 en *Computational Linguistics* 27 (1), marzo

BALLESTER, A. / A. MARTÍN MUNICIO / F. PARDOS / J. PORTA / R. RUIZ UREÑA / F. SÁNCHEZ LEÓN (2002) «Combining Statistics on *n*-grams for Automatic Term Recognition», en *Actas del III Congreso de LREC, Language Resources and Evaluation Conference*, de la Asociación Europea de Recursos Lingüísticos (ELRA), 29-31 de mayo de 2002, Las Palmas

REAL ACADEMIA DE CIENCIAS EXACTAS, FÍSICAS Y NATURALES (1999) *Diccionario Esencial de las Ciencias,* Espasa-Calpe, Madrid

XCES: Corpus Encoding Standard for XML, Document XCES 0.2 http://www.xml-ces.org