Actas del XXIV Congreso

Aprendizaje de lenguas, uso del lenguaje y modelación cognitiva: perspectivas aplicadas entre disciplinas

Madrid, UNED 2006
ISBN 978-84-611 6897 2

ETIQUETADO DEL CORPUS TXTCERAM ORIENTADO A LA EXTRACCIÓN DE INFORMACIÓN CONCEPTUAL

AMPARO ALCINA CAUDET
ANNA ESTELLÉS PALANCA
Universitat Jaume I

RESUMEN

En el presente trabajo nos centramos en los corpus como recurso para la investigación en procesamiento del lenguaje natural con fines terminológicos. Basándonos en la guía TEI, presentamos las plantillas utilizadas para etiquetas nuestro corpus TxtCeram y sus características para trabajar con WordSmith, una herramienta de análisis textual. Presentamos un experimento que estudia la frecuencia de hiperónimos en las secciones de introducción, para probar la adecuación de la herramienta WordSmith trabajando con carpus etiquetado. Palabras clave: Terminografía, lingüística de corpus, etiquetado de corpus, análisis textual, corpus TxtCeram.

ABSTRACT

In this paper we will focus on corpora as a resource for researching language processing for terminological purposes. Based on the TEI guide, we present the templates used to tag our TxtCeram corpus and its features when working with WordSmith, a text analysis tool. We present an experiment for studying the frequency of hyperonyms in the introduction section of texts, while testing WordSmith's suitability to work with our tagged corpus. Keywords: Termonography, corpus linguistics, corpus tagging, análisis textual, TxtCeram corpus.

1. INTRODUCCIÓN

El *uso de corpus* viene siendo un recurso muy valioso en todas las áreas que tratan con el procesamiento del lenguaje natural, donde se incluye el uso de corpus por parte de lingüistas para llevar a cabo sus estudios o investigar en *Terminología*. En este trabajo, nos centramos en el corpus como recurso para investigar el procesamiento de lenguaje natural con fines terminológicos. En Terminología, los investigadores se centran en lenguajes de un dominio específico para investigar aspectos terminológicos y por tanto, los corpus representan un modelo de lenguaje realista para su estudio (Bowker 1996). Es decir, la necesidad de la investigación lingüística basada en el uso de corpus se debe a la necesidad de estudiar el lenguaje a través de ejemplos reales (Sinclair 1991). Tognini-Bonelli (1996) explica la diversidad de aplicaciones de la *lingüística de corpus* y ofrece una amplia descripción sobre sus usos y objetivos. La autora defiende el uso de corpus como la base de su modelo empírico para investigación, que consiste en: la observación de hechos lingüísticos, la formulación de hipótesis y generalizaciones basada en patrones de datos y la subsiguiente derivación de observaciones teóricas (Tognini-Bonelli 2001). Además, la existencia de herramientas de análisis textual, *parsers*, así como códigos o estándares para codificar textos, con el fin de hacer que los textos sean procesables automáticamente, ha potenciado el abanico de estudios basados en corpus (véase Garside, Leech y McEnery 1997). Autores como Ahmad y Rogers (1997) explican que el tratamiento del corpus dependerá de criterios explícitos. Básicamente, el corpus crudo se utiliza para calcular la frecuencia de palabras o fragmentos, entre otros. Por otro lado, un corpus etiquetado puede, por ejemplo, estar analizado lingüísticamente o ser una versión anotada del corpus crudo que contenga no sólo las palabras que lo constituyen

sino información sobre la situación comunicativa (Ahmad y Rogers 1997). Además, entre otros, el *etiquetado* puede orientarse a aspectos semánticos, morfosintácticos, contextuales y/o estructurales, por tanto, el objetivo de la compilación determinará la información que se etiquetará (Alcina 2001). Algunos lenguajes informáticos permiten el etiquetado textual y su procesamiento automático: TEI y DocBook son dos ejemplos. Actualmente, TEI (*Text Encoding Initiative*) proporciona unas guías, denominadas *TEI Guidelines*, que constituyen un estándar internacional y interdisciplinario que facilita a bibliotecas, museos, publicaciones, estudiantes o personas interesadas, la representación de una variedad de textos lingüísticos y literarios orientada a la investigación, docencia o conservación digital (Burnard y Sperberg-McQueen 2002). DocBook también proporciona un sistema para escribir documentos estructurados basado en SGML o XML (Walsh y Mueliner 1999).

Para analizar los corpus etiquetados, se ha desarrollado una amplia gama de herramientas y recursos que pueden ayudar a la investigación terminológica en terminografía. Por una parte, los procesos como la adquisición manual de datos terminológicos a partir de texto se están sustituyendo por técnicas automáticas que ayudan a realizar esta dura tarea. Por otra, las bases de conocimiento y sistemas de ontologías se están viendo como potentes herramientas para gestionar datos terminológicos y conocimiento (Millar y Fellbaum 1991; Guarino 1995; Gruber 1993).

En este trabajo presentamos el diseño del corpus TxtCeram y el modelo de etiquetado que hemos adoptado. El etiquetado de corpus será el primer paso para la extracción de datos con el fin de construir una ontología terminológica, y se enfoca como objeto de análisis e investigación para prácticas metodológicas hacia la construcción de una ontología terminológica. Finalmente, presentamos algunas metodologías para mejorar el procesamiento del corpus etiquetado cuando se utiliza en una herramienta de análisis textual como WordSmith.

2. Proyecto TxtCeram

TxtCeram es un proyecto de investigación que está siendo desarrollado por el grupo de investigación Tecnolettra (Universidad Jaime I, http://tecnolettra.uji.es). El proyecto se centra en la extracción semiautomática y el análisis conceptual formal de términos de la cerámica. Su principal objetivo es probar la eficacia de algunas de las herramientas informáticas que se utilizan para diseñar un sistema integral de terminología asistida y los beneficios del uso de este sistema en la mediación lingüística.

Para ello, se ha creado un *corpus* de *lenguaje especializado* del campo de la cerámica. Actualmente, contamos con un corpus de 12,6 MB (formato txt) en español y hemos empezado a compilar el corpus inglés, que a día de hoy cuenta con 1,16 MB (formato txt); esto significa 2,8 millones de palabras en el corpus de español y sobre unas 250.000 palabras en el inglés. Los textos que se han incluido son obras originales que pertenecen al campo de la cerámica. El proceso de compilación del corpus se ha realizado bajo criterios exhaustivos de documentación y evaluación basados en procedimientos de búsqueda cualitativa (Alcina 2005). Nos hemos centrado en recoger aquellos libros que contenían no solo información terminológica sino contextos relevantes y explicaciones sobre los procesos y los elementos relacionados con la cerámica. La selección incluye once libros que tratan la cerámica y incluyen datos sobre: productos y tecnologías para esmalte y coloración, fabricación, usos de la baldosa, estructuras, compuestos químicos, materiales, procedimientos y procesos.

Resumiendo, nuestro corpus reúne las siguientes características: es un corpus escrito, sincrónico (las obras se han publicado durante el mismo periodo, 1980-1997), especializado, textual (obras completas), documentado (incluye no solo texto sino también información

contextual) y multilingüe (español, inglés y catalán). Para más información sobre tipología de corpus véase Sinclair (1991), Tognini-Bonelli (1996 y 2001), McEnery y Wilson (1996), Garside, Leech y McEnery (1997), Biber, Conrad y Reppen (1998) o Sánchez-Gijón (2004).

3. HTML, TEI P4 LITE, CES & DOCBOOK

Si nos centramos en los elementos del etiquetado, se pueden distinguir diferentes ramas según el objetivo del proceso de etiquetado. A parte del uso o no uso de etiquetadores automáticos, el etiquetado se puede orientado principalmente a aspectos morfológicos, sintácticos, léxicos, semánticos o/y discursivos. En nuestro trabajo, exploramos el etiquetado de la información macro-estructural y contextual del corpus. Estos datos servirán como base para pruebas de investigación, que incluyen el diseño de una ontología terminológica. La macro-estructura de los textos proporciona información sobre el género textual y las partes que constituyen el documento. Puede resultar útil en la extracción de información terminológica, ya que la frecuencia de términos y su posición puede depender del foco de investigación. La información contextual será útil para entender las características del texto, y por tanto, el entorno del término.

Como hemos explicado anteriormente, TEI Lite y DocBook son dos lenguajes para etiquetar textos. Son muy similares ya que las últimas versiones de ambos están basadas en XML y complementados con un DTD (*document type description*) que será más o menos complejo según las necesidades del usurario. Por otra parte, encontramos HTML, el lenguaje estándar para estructurar información en *Internet*. Si analizamos las características especiales de cada uno de estos lenguajes, observamos que aunque HTML se pensó como un modo de combinar y crear recursos digitales, su arquitectura no soporta información semántica. DocBook se concibe como un DTD para la publicación de obras, su arquitectura se centra en permitir tanto la representación electrónica como la impresa, sin que sus productos resulten diferentes uno de otro, la información semántica que aportará irá ligada a este aspecto de representación. CES (*Corpus Encoding Standard*) y TEI cumplen el mismo estándar y están basados en las guías del grupo EAGLES, la principal diferencia entre ambos es el nombre que reciben algunas etiquetas cuya función es la misma. Ambos se crearon para representar documentos ya existentes (Rahtz, Walsh y Burnard 2004) y han servido de base para etiquetar el *British Nacional Corpus*. Están diseñados para usarse en investigación en ingeniería del lenguaje y en aplicaciones basadas en lenguaje natural, como lenguaje de etiquetado de información estructural y como un medio para añadir información semántica a los textos. La última versión de CES define bastantes elementos que profundizan en niveles de detalle de descripción del texto (como, por ejemplo, marcado de oraciones o palabras clave dentro de oraciones); TEI Lite, en cambio, se limita a un análisis más superficial (estructura y situación comunicativa). En lo referente a estructura e información semántica de alto nivel ambos comparten las mismas etiquetas. En nuestro trabajo hemos partido de la versión TEI P4 en su edición simplificada TEI P4 Lite ya que basta para definir la información que marcaremos, su arquitectura es sencilla, está documentado y proporciona grandes facilidades para diseñar la DTD (la herramienta TEI PizzaChef, http://www.tei-c.org/pizza.html).

4. ETIQUETADO

Para sistematizar el proceso de etiquetado hemos elaborado una plantilla para la *información contextual* y dos plantillas para la *macro-estructura*: una para etiquetar los libros completos y

capítulos, y la otra para etiquetar artículos. La plantilla de información contextual se aplicará tanto en libros como artículos y la aplicación de las plantillas de información macro-textual aportará elementos descriptivos que dependen de la naturaleza del documento, como veremos a continuación.

4. 1. Información Contextual

La información contextual se distribuye en seis grupos donde se incluyen las etiquetas, estos grupos pertenecen al encabezado del documento y se incluirán dentro de la etiqueta <head> del documento xml, son los siguientes (tabla 1):

	Etiqueta	Función
Fichero	<fileDesc>	Contiene etiquetas de descripción del fichero
	<extent>	Tamaño en Kb
	<idno>	Código de identificación del fichero
Documento electrónico	<publicationStmt>	Creación del documento electrónico
	<publisher>	Publicación digital
	<distributor>	Distribuidor digital
	<pubPlace>	Lugar de publicación
	<date>	Fecha de creación
Obra original	<sourceDesc>	Contiene etiquetas de información bibliográfica
	<biblStruc>	Información bibliográfica estructurada
	<analytic>	Información sobre el texto (parte específica, i.e. artículo, capítulo)
	<author>	Autor si es diferente al de la obra completa
	<title>	Título
	<monogr>	Información sobre la obra a la que pertenece el documento
	<editor>	Editor
	<imprint>	Información sobre su edición impresa
	<pubPlace>	Lugar de publicación
	<publisher>	Responsable de su publicación
	<date>	Fecha de publicación
	<idno>	ISBN o ISSN
	<biblScope type="pages">	Número de páginas del original
	<title>	Título obra
	<author>	Autor/es obra
	<extent>	Número de páginas
Etiquetado	<encodingDesc>	Descripción del lenguaje de etiquetado (por ejemplo, xml-TEI)
	<projectDesc>	Descripción del proyecto de etiquetado
	<editorialDecl>	Aspectos legales, uso y distribución
Tipo de documento	<profileDesc>	Contiene etiquetas para tipo de documento, descripción de la situación comunicativa
	<creation>	Formato original: papel, audio, etc.
	<langUsage>	Tipo de lenguaje
	<language>	Idioma/s
	<textClass>	Tema, según la Biblioteca del Congreso.
	<keywords>	Palabras clave. Puede ir listadas mediante etiquetas <list> y <item>
	<classCode>	Código que corresponde al tema de <textClass>
	<catRef>	Sistema de codificación usado para <classCode> y <textClass>. Biblioteca del Congreso.
	<textDesc>	Descripción breve del texto
	<partieDesc>	Descripción del perfil del emisor y receptor
	<settingDesc>	Modo: escrito, oral, etc.
Revisiones	<revisionDesc>	Contiene etiquetas para la relación de eventos sucedidos durante el proceso de etiquetado o creación
	<change>	Cambios
	<date>	Fecha de los cambios
	<respStmt>	Responsable de los cambios
	<item>	Segmento modificado y segmento original

Tabla 1. Información Contextual

4.2 Macro-estructura

Para la macro-estructura, hemos simplificado el catálogo de etiquetas y hemos elaborado dos esquemas que coinciden en todas las etiquetas excepto en una (y). Esta descripción se ha fundamentado, principalmente, en teorías basadas en el género textual (véase Swales 1990; Bahtia 2002). Estos autores defienden la importancia de los propósitos de la comunidad de hablantes de un lenguaje especializado, sugiriendo que un lenguaje especializado se caracteriza por un conjunto de propósitos comunicativos acordados por los miembros de la comunidad discursiva.

Los elementos de la macro-estructura se incluyen dentro de las etiquetas <text> y varían según el género textual de la obra original. Las variaciones de los atributos de la etiqueta , descritas en la DTD, permiten la representación de las diferencias existentes entre la macro-estructura de un libro, un capítulo o un artículo. La Tabla 2 agrupa las etiquetas que se utilizan para marcar un artículo y las utilizadas para libros, los capítulos también tomarán etiquetas del esquema. Como observamos, la estrategia que se ha seguido se basa en permitir una comparación inter-textual de las secciones de los textos del corpus.

Etiquetas	Función	Ruta
<front>	Información preliminar	<text>
	Sección resumen (artículos)	<text>, <front>
	Sección prefacio (libros)	<text>, <front>
	Sección índice	<text>, <front>
<body>	Sección cuerpo	<text>
	Introducción	<text>, <body>
	Conclusión	<text>, <body>
<head>	Título	<text>, ...
<back>	Agradecimientos, apéndice y bibliografía	<text>
	Apéndices o tablas añadidas al final del documento	<text>, <back>
	Bibliografía	<text> <back>

Tabla 2. Etiquetado macro-estructural

5. PROCESAMIENTO

La extracción de candidatos a término automática o semiautomática se considera un factor acelerador del proceso de investigación para la creación de recursos profesionales a largo plazo; como es nuestro caso y como ha ocurrido en otros trabajos (Reimerick 2002; Pérez Hernández 2002; Faber 2002). En este apartado, presentamos la herramienta de análisis que hemos utilizado para sacar provecho de la información macro-estructural y contextual documentada en los textos y el experimento que hemos realizado para probar su eficacia en el análisis de secciones específicas.

5. 1 La herramienta WordSmith

WordSmith es una herramienta de análisis textual que trabaja con texto crudo o texto etiquetado sencillo. Esta herramienta, desarrollada por Mike Scott (*Oxford University*), resulta útil en ingeniería lingüística; un ejemplo de su uso es su aplicación basada en el concepto de género textual (Alcina 2005). WordSmith (en su versión 3) integra tres herramientas: WordList – Proporciona listas de palabras o de grupos de palabras a partir de un texto ordenadas alfabéticamente o por frecuencia de aparición. Concord – Proporciona listas de palabras y sus contextos. KeyWords – Permite la extracción de palabras clave basada en un método estadístico.Estas herramientas permiten el uso de *stop list* y ficheros de

lematización. Una *stop list* es una lista de palabras que no queremos incluir en los cálculos o procesos de análisis textual. Un fichero de lematización es una lista de palabras agrupadas por un mismo lema que permitirá el cálculo o análisis de todas las palabras de un mismo lema agrupadas como una misma entrada. La configuración de esta herramienta para utilizar la información etiquetada dependerá del objetivo de la investigación.

5.2 Nuestro experimento

Hemos configurado la herramienta para proporcionar una lista de frecuencia de la sección de introducción de los textos y compararla con una lista de frecuencia de la sección cuerpo entera. La prueba se ha basado en 16 ficheros que conforman 4 de los 11 libros que contiene el corpus TxtCeram de español. Cada capítulo de cada libro está etiquetado según los esquemas mostrados anteriormente así como cada libro. El material consiste en 306.068 palabras de texto etiquetado y se ha utilizado WordList para realizar el análisis estadístico.

Una vez los ficheros han sido seleccionados, hemos aplicado una *stop list* (Figura 2) y un fichero de lemas (Figura 3). El proyecto TxtCeram ha compilado una *stop list* para el corpus español que incluye pronombres, artículos, preposiciones, numerales, relativos y adverbios comunes. El fichero de lematización de TxtCeram incluye una lista de 5548 verbos españoles, una compilación de verbos desarrollada a partir de la colección realizada por Alonso (Alonso 1989).

Hemos diseñado un fichero de etiquetas que incluye las etiquetas que deberán tenerse en cuenta de modo que se ignoren las demás (Figura 4). Esto eliminará ruido en el corpus y no se contarán etiquetas como las del grupo de información contextual (<autor>, <analytic>, etc.), que resultan irrelevantes en nuestro experimento. Nuestro fichero de etiquetas contiene las etiquetas para identificar la sección de introducción ( ) y la sección de cuerpo (<body>). Una vez creado el fichero, hemos configurado la herramienta para cargar el fichero de etiquetas (Figura 5). El siguiente paso ha sido seleccionar el apartado que queríamos usar para realizar el cálculo de la lista de frecuencia. Primero hemos calculado una lista de frecuencia de palabras en la sección de introducción (Figura 6) y a continuación, del cuerpo, donde hemos sustituido las etiquetas y por <body> y </body>. Los resultados (Figura 7 y Figura 8) mostraron, en una primera impresión, que la lista de frecuencia de palabras de las secciones de introducción presenta hiperónimos en las primeras posiciones y que estos comparten un alto grado de frecuencia. Comparando la sección de cuerpo, distinguimos que hay una menos frecuencia de hiperónimos y su tasa es también menor. En la lista de la secciones de introducción, los contextos de las palabras han mostrado que las 20 primeras de la lista no van seguidas de descriptores (propiedades) o hipónimos. En cambio, en la sección de cuerpo aparecen descriptores cerca de los primeras 10 palabras calculadas en la lista de frecuenta y a partir de la posición 11ª ya empezamos a observar que sus contextos contienen muchos términos específicos (como tipos de bizcocho o de esmalte o términos como *rodillo*). En principio, a la espera de una investigación más exhaustiva, podemos concluir que las secciones de introducción en las obras de cerámica no aparecerán descriptores o propiedades pero sí hiperónimos, que puede resultar útil para extraer los primeros niveles de un árbol terminológico sobre cerámica.

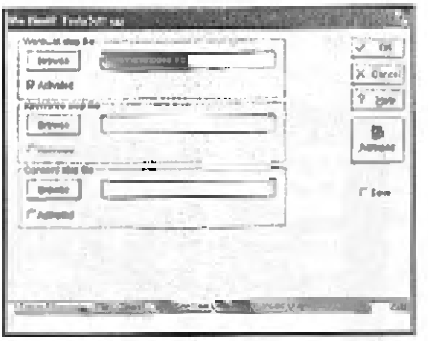

Figura 2. *Stop list*

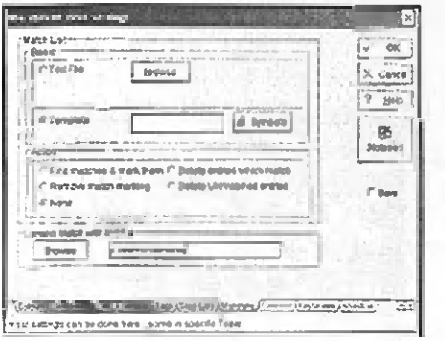

Figura 3. Lematización

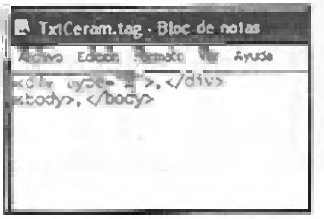

Figura 4. Lista de etiquetas

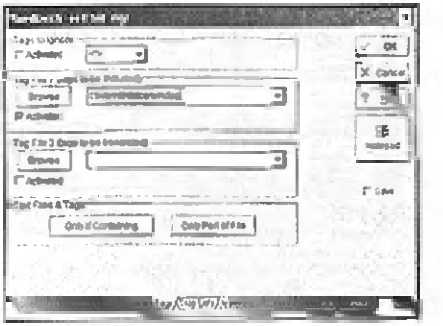

Figura 5. Configuración lista de etiquetas

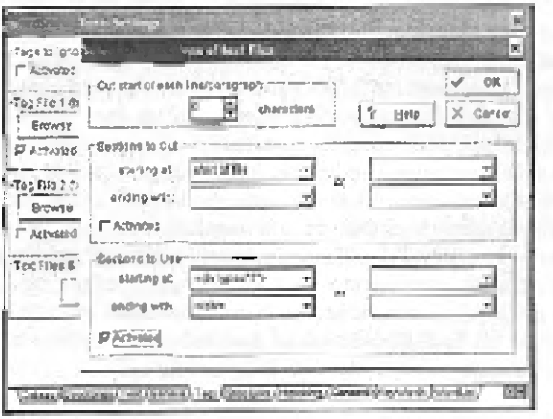

Figura 6. Selección secciones introducción

Figura 7. Resultados secciones introducción

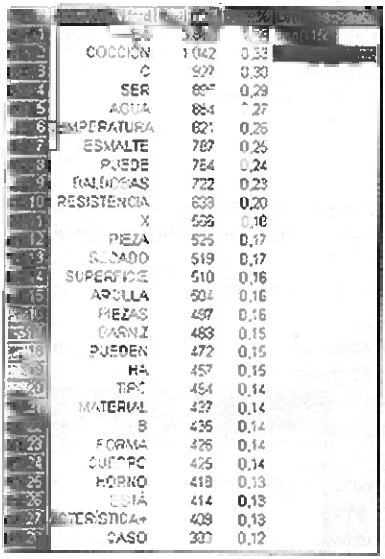

Figura 8. Resultados secciones cuerpo

6. Conclusión

A partir de nuestro experimento se prueba que los esquemas de etiquetado presentados aquí funcionan según lo esperado, lo que prueba que su diseño. Funcionan correctamente con el analizador textual WordSmith y, por tanto, mejoran el proceso de extracción de candidatos a término. El diseño de etiquetado ha demostrado flexibilidad para trabajar con diferentes secciones, ya que se permite la selección específica de apartados y la combinación de búsquedas entre distintos apartados. Esto puede resultar muy interesante en investigación sobre posiciones de términos en texto estructurado. Las secciones no relevantes pueden evitarse y se permite una investigación más focalizada en aquellas secciones que sí resulten relevantes al objeto de estudio. Además, este modelo de etiquetado permite gestionar y controlar la información contextual, lo que resulta un aspecto interesante para aquellos estudios terminológicos basados en datos relacionados con la situación comunicativa. La herramienta WordSmith ha mostrado un comportamiento eficaz en el análisis, si bien no descartamos la posibilidad de utilizar una herramienta más robusta capaz de procesar los datos especificados en la DTD, de modo que podamos beneficiarnos de la información semántica que contienen.

7. BIBLIOGRAFÍA

Ahmad, K. & Rogers M. 1997. Corpus Linguistics and Terminology Extraction. In S. E. Wright, G. Budin (eds.), *Handbook of Terminology Management*. Amsterdam & Philadelphia: John Benjamins, vol 2, pp. 725-760.

Alcina, A. 2001. Automatización de Tareas en la Elaboración de Diccionarios Terminológicos. In *Proceedings of Terminologia i documentació. I Jornada de Terminologia i Documentació*. Barcelona: Universitat Pompeu Fabra pp. 51-60.

Alcina, A. 2005. La Implementación del Concepto de Género Textual en los Corpus Electrónicos para Traductores. In García Izquierdo, I (ed), *El género textual y la traducción*. Berna: Peter Lang, pp. 93-114.

Alcina, A.; Soler, V. & Estellés, A. 2005. Internet como Instrumento para la Documentación en Terminología y Traducción. In Sales Salvador, D. (ed), *Documentarse para Traducir*. Granada: Comares.

Alonso Moro, I. 1989. *Verbos Españoles*. Madrid: Difusón S.L.

Bhatia, V. K. 1993. *Analysing Genre. Language Use in Professional Settings*. London: Longman.

Biber, D.; Conrad S. & Reppen R. (eds.) 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bowker, L. (1996). Towards a Corpus-based Approach to Terminography. *Terminology*, 3(1) pp. 27-52.

Burnard, L. & Sperberg-McQueen C. M. 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Chalottersville, Bergen: Text Encoding Initiative Consortium.

Faber, P. & Jiménez, C. (eds.) 2002. *Investigar en Terminología*. Granada: Comares.

Garside, R.; Leech G. and McEnery A. (Eds.) 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), pp. 199-220.

Guarino, N. 1995. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies*. special issue, 43(5-6), pp. 625-640.

Martin, L.E. 1990. Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 252-262.

McEnery, T. & A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Miller, G. A. & Fellbaum C. 1991. Semantic Networks of English. In *Cognition*, special issue, 197-229. Reprinted in Levin B. and Pinker, S. (eds.) *Lexical and Conceptual Semantics*. Cambridge, MA: Blackwell, pp. 197-229.

Pérez Hernández, C. 2002. Terminografía Basada en Corpus. In Faber, P. & Jiménez, C. (eds.) *Investigar en Terminología*. Granada: Comares.

Rahtz, S.; Walsh, N. & Burnard, L. 2004. A Unified Model for Text Markup: TEI, Docbook, and beyond. In *Proceedings of XML Europe 2004*. Amsterdam : DeenIX (digital edition).

Reimerink, A. 2002. El Análisis de Corpus para un Fin Práctico: Tendencias en el Uso de los Verbos en la Redacción de Artículos de Investigación. In Faber P. & Jiménez, C. (eds.) *Investigar en Terminología*. Granada: Comares.

Sánchez-Gijón, P. 2004. *L'ús de corpus en la traducció especialitzada*. Barcelona: IULA, Universitat Pompeu Fabra.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Swales, J. M. 1990. *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.
Tognini-Bonelli, E. 1996. *Corpus Theory and Practice*. Birmingham: TWC.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins.
Walsh, N. & Muellner, L. 1999. *DocBook: The Definitive Guide*. O'Reilly & Associates, Inc.