Faber, Pamela, Clara Inés López Rodríguez y María Isabel Tercedor (2001) "La utilización de técnicas de corpus en la representación del conocimiento médico". *Terminology* 7 (2), 167-197. Ámsterdam/Filadelfia: John Benjamins.

UTILIZACIÓN DE TÉCNICAS DE CORPUS EN LA REPRESENTACIÓN DEL CONOCIMIENTO MÉDICO[1]

Pamela Faber, Clara Inés López Rodríguez y María Isabel Tercedor Sánchez
Facultad de Traducción e Interpretación
Universidad de Granada
C/Puentezuelas 55
18071 Granada
Spain

Abstract:

Advances in corpus linguistics are of vital importance in terminology. The information obtained from corpora can be used to complement data already codified in dictionaries and termbases. In this article, we describe a framework of linguistic analysis that facilitates the extraction of conceptual information from corpora, and thus contributes to the study and analysis of terminological contexts. We are presently using this methodology in a research project called ONCOTERM. One of the objectives of this project is to elaborate a bilingual terminological database, whose conceptual structure is an extension of an already existing knowledge resource, the Mikrokosmos Ontology.

In our termbase, medical concepts are organized in categories represented by templates, which are systematically applied to all category members. The application of the template to more specific concepts generates values that show the inheritance of knowledge structures within a specialized domain. The definitional information within each term entry is thus totally coherent with the information regarding other terms within the same conceptual category. This is conducive to the specification of a language of terminographic definition, which is concise, consistent and applicable not only to the domain of oncology, but also extensive to other medical domains and other languages.

Key words: medical terminology, corpus studies, knowledge representation, category template, conceptual structure, ontology

1. Introducción

La *lingüística de corpus* tiene una importante repercusión en el estudio del lenguaje y en el desarrollo de aplicaciones informáticas de utilidad para el procesamiento del lenguaje natural, la lexicografía, el análisis y la etiquetación sintáctica de textos, la traducción automática, las memorias de traducción y la gestión de terminología. Estas áreas han adoptado una metodología más empírica gracias a la explotación de corpora, es decir, colecciones de textos en soporte electrónico de un tamaño limitado, recopilados para representar una variedad lingüística concreta (McEnery y Wilson 1996: 24).

Desde el final de la década de los ochenta, la terminología viene extrayendo de textos especializados el conocimiento terminológico necesario para la creación de bases de datos terminológicas. El análisis textual que brindan los corpora ha contribuido a la consolidación de una terminología descriptiva aplicable a sistemas de recuperación de información y a la traducción. Han proliferado estudios cuyo objetivo es la extracción de términos y la adquisición de conocimiento especializado. Entre los destinados a la detección y extracción automática de términos, podemos citar Lauriston (1993), Bourigaut (1994), Daille (1994), Jacquemin y Royauté (1994) y L'Homme (1996), donde se explicitan los patrones morfosintácticos de las unidades terminológicas, al igual que se había hecho en lexicografía monolingüe (Benson et al. 1986). Algunas investigaciones se centran en el dominio de la medicina, como ocurre con Jacquemin y Royauté (1994), L'Homme (1999) y Ahmad, Davies, Fulford y Rogers (1994).

El reconocimiento de patrones sintácticos recurrentes facilita el análisis conceptual que llevan a cabo traductores, terminógrafos y expertos en ingeniería del conocimiento, por lo que constituyen *sondas* que permiten extraer de textos conocimiento sobre un determinado dominio (Ahmad y Fulford *apud* Meyer y Mackintosh 1996: 21). A partir de estos patrones de formación de términos, se puede recuperar el conocimiento necesario para la indexación de tesauros médicos (Jacquemin 1997) y para la formulación de definiciones terminográficas (Pearson 1998). Al fin y al cabo, como afirman Borigaut y Slodzian (1999: 31), la terminologización es un proceso paralelo a la elaboración conceptual.

En cualquier caso, todos estos estudios basados en corpus ponen de manifiesto que, al hacer explícito el contexto del término, sale a relucir información sobre su significado y sobre su uso (Pearson 1998: 191). Según L'Homme et al. (1999: 32-33), el contexto ayuda a recuperar términos asociados, elementos definitorios, relaciones de

sinonimia y relaciones explícitas entre conceptos vinculados con ellos semánticamente; asimismo, a partir del contexto es posible verificar la información conceptual y gramatical extraída de obras de referencia, especificar el significado de siglas y añadir información enciclopédica no incluida en la definición. Para esto, no basta con el manejo de programas de análisis léxico que elaboran listas de frecuencia, preprocesan el texto antes del análisis u obtienen de forma automática líneas de concordancia de un término. Es preciso interpretar los datos dentro de un marco coherente y sistemático de análisis lingüístico de base textual (Bourigaut y Slodzian 1999: 29; L'Homme et al. 35).

En efecto, tanto el terminólogo como el traductor de textos especializados precisan de una metodología para la extracción de información a partir de corpora, información que deberá ser complementada con los datos extraídos de diccionarios y de bases de datos terminológicas ya existentes. Esta metodología será válida si pone de manifiesto el carácter comunicativo (Cabré 1999) y socio-cognitivo (Temmerman 2000) de los términos.

En este artículo se propone un marco de análisis lingüístico para extraer información conceptual de corpora de textos especializados y así facilitar el estudio de contextos terminológicos y de la interrelación entre la lengua general y los lenguajes especializados. La aplicación de dicho marco se está llevando a cabo en ONCOTERM (PB98-1342), un proyecto de investigación interdisciplinar sobre terminología médica, financiado por el Ministerio de Educación. El objetivo principal de ONCOTERM es la elaboración de un sistema de información sobre el subdominio biomédico de la ONCOLOGÍA. Teniendo en cuenta que este sistema se deriva de la creación y configuración de una amplia base de datos terminológica sobre el cáncer, el proyecto se sustenta también sobre los siguientes objetivos operativos:

- Crear un corpus de textos médicos tanto en español como en inglés.
- Especificar un lenguaje de definición terminográfica conciso, consistente y aplicable no sólo al subdominio de la oncología, sino también a otras especialidades médicas y a otras lenguas.
- Elaborar un inventario de relaciones conceptuales específicas del EVENTO MÉDICO (véase 3.1.) y, en particular, del EVENTO MÉDICO ONCOLÓGICO.
- Configurar una base de datos terminológica articulada en torno a la estructura hallada en la definición de los términos.

Uno de los resultados principales de este proyecto es la especificación de lo que podría llamarse *esquemas categoriales* (category templates) que refleje la organización interna característica de cada categoría de conocimiento médico. Esta configuración se basa en la información extraída de textos con la utilización de técnicas propias de la lingüística de corpus, y validada por un especialista (Bourigaut y Slodzian 1999: 29-30). Este *esquema categorial* asegura la consistencia tanto de las definiciones terminográficas como de la ontología y la base de datos terminológica sobre el cáncer que estamos elaborando.

2. Metodología para la extracción de conocimiento especializado en ONCOTERM

2.1. ONCOTERM

La parte fundamental de ONCOTERM es la representación de la estructura conceptual del dominio de la ONCOLOGÍA MÉDICA, sus conceptos pertinentes, características e interrelaciones. La estructuración conceptual inicial del dominio se consigue mediante la elaboración de jerarquías terminográficas, basadas en la extracción de información conceptual de textos especializados y diccionarios médicos.

La metodología de establecer jerarquías mediante el análisis de definiciones lexicográficas se remonta a Amsler (1980), quien deriva información hiponímica en relación a los sustantivos del lenguaje general. La premisa básica es que la información contenida en diccionarios constituye una red léxico-conceptual que necesariamente tiene correspondencia con el conocimiento expresado. Posteriormente, Meijs y Vossen (1992: 144-145) sostendrán que estas cadenas léxicas revelan los parámetros de conocimiento especificados en las *differentiae* de los lexemas así relacionados, una idea que también está presente en las jerarquías léxicas de *WordNet*, como se aprecia en el ejemplo:

(1) {robin, redbreast}@ →{bird}@ → {animal, animate_being}

(Miller 1998: 25)

Este principio metodológico es aplicable también a la terminología, aunque, en este caso, las jerarquías de sustantivos terminológicos son más largas, ya que empiezan a niveles mucho más específicos. En cualquier caso, ambas cadenas acaban con genéricos pertenecientes a la lengua general.

(2) {electron beam radiation therapy}@→{external radiation therapy}@ →{radiation therapy}@ →{treatment}@ →{event}@→{ALL}

Como puede percibirse, términos como *electron beam radiation therapy* y *external radiation therapy* están en una relación de subordinación con *treatment* y *event*. Esto facilita la ubicación del campo especializado dentro de un contexto más amplio, que también incluye la representación de lo que podría llamarse *conocimiento sobre el conocimiento*, es decir, las diferentes relaciones conceptuales que vinculan los conceptos entre sí. Como queda ilustrado en (3), dichas relaciones conceptuales pueden extraerse de información textual en la forma de concordancias. Aquí se pone en juego el segundo procedimiento de análisis:

(3) Parametrización del conocimiento pertinente al concepto RADIATION THERAPY

RADIATION_THERAPY
RADIATION_SOURCE_LOCATION 1 516-520, 1981. Bagshaw MA: External **radiation therapy** of carcinoma of prosta 2 r near the tumor. Also called internal **radiation therapy** or implant radiation.
↳ **INTERNAL_RADIATION_THERAPY** 3] 4. Laser therapy or interstitial **radiation therapy** for endobronchial 4 ntimeters thick: 1. Intracavitary **radiation therapy**. In most instances, 6,
↳ **EXTERNAL_RADIATION_THERAPY** ↳ **BEAM_TRAJECTORY** 5 c radiosurgery and stereotactic **radiation therapy**. stereotaxis (
↳ **BEAM_TYPE** 6 il and intraoperative electron beam **radiation therapy** on the outcome of pati 7 radiosensitizers, or particle-beam **radiation therapy**. [14-17] 4. Isotre 8 to chemotherapy.[7-9] Fast neutron beam **radiation therapy** or accelerated hyper 9 ltileaf collimator.[25] Proton-beam **radiation therapy** is also under investig 10 al with mixed-beam (neutron/photon) **radiation therapy**, compared to standard
RADIATION_BEAM_TARGET ↳ **BODY_PART** 11 m the NWTS-3 demonstrate that abdominal **radiation therapy** does not provide sig 12 metastases are identified, whole brain **radiation therapy** (30 gray in 2 gray f 13 s incorporating chemotherapy plus chest **radiation therapy** are listed below. Th 14 pts to lower the dose of craniospinal **radiation therapy** to 2,340 cGy have resu 15 tive chest wall and regional lymph node **radiation therapy** are undergoing reass 16 Jett JR, McGinnis WL, et al.: Thoracic **radiation therapy** alone compared wit
↳ **FIELD_COVERAGE** 17 of whole pelvis versus small-field **radiation therapy** for carcinoma of prost 18 us 2 months of ABVD plus extended-field **radiation therapy** is being conducted b 19 arbell NJ, Silver B, et al.: Wide-field **radiation therapy** with or without ch
RADIATION_ADMINISTRATION

↳ **DOSAGE**	
20	ged remissions. The need for low-dose **radiation therapy** is under study.[1] The
21	pse after initial wide-field, high-dose **radiation therapy** have a good prognosi
22	using accelerated superfractionated **radiation therapy** for advanced squamous
23	e.[6,11-18] 3. Novel fractionation **radiation therapy** clinical trials are un
24	l evaluation: 1. Hyperfractionated **radiation therapy** to improve tumor contr
↳ **INTENSITY**	
25	ned 5-fluorouracil and supervoltage **radiation therapy** of locally unresectabl
26	years' experience with megavoltage **radiation therapy**. Cancer 37(6): 2605-26
27	ectron-beam irradiation or orthovoltage **radiation therapy** may be used to pal

Investigaciones previas han explotado las líneas de concordancia a la hora de identificar relaciones léxicas básicas como la sinonimia, la paráfrasis o la sustitución (Pearson 1998) o funciones léxicas fundamentadas en la *Meaning Text Theory* (Mel'cuk 1996, Mel'cuk et al. 1984). Se ha demostrado que las funciones léxicas de Mel'cuk hacen sistemático, en el campo de la terminología médica, el estudio de las relaciones semánticas entre los elementos de las unidades terminológicas (Laporte y L'Homme 1996), la determinación del radio colocacional de los términos (Tercedor 1999, Faber y Tercedor 2001)[2] y el reconocimiento de la cohesión léxica de los textos (López Rodríguez 2001). No obstante, para que los recursos terminológicos desarrollados sean exhaustivos y reutilizables en el campo del procesamiento del lenguaje natural, no habría que quedarse en un inventario cerrado de funciones léxicas, sino que habría que refinar aún más las relaciones conceptuales activadas en los textos. Estas relaciones, como se verá en el apartado 3, sirven de vínculos entre conceptos y permiten un modelado conceptual y terminológico caracterizado por la coherencia y la concisión.

En consecuencia, en nuestro proyecto, las concordancias proporcionan un esquema básico de relaciones conceptuales sobre el cual puede modelarse el resto del subdominio. Las concordancias en (3) ilustran que, en la estructuración del subdominio RADIATION THERAPY, hay que tener en cuenta la diferencia entre 'internal radiation therapy' y 'external radiation therapy', y que tanto *intracavitary radiation therapy* como *interstitial radiation therapy* son tipos de radioterapia interna. Sin embargo, la radioterapia externa, que es el tipo de radioterapia más común, tiene un abanico de distinciones conceptuales más amplio, como el tipo de radiación utilizada, la trayectoria del haz, la parte del cuerpo enfocada, la extensión del campo enfocado, así como la dosis e intensidad de la radiación.

Por otra parte, el corpus también muestra la existencia de otros marcos de conceptualización complementarios para RADIATION THERAPY, que están en relación con su estatus de EVENT[3]. Estos marcos tienen que ver con factores contextuales, como, por

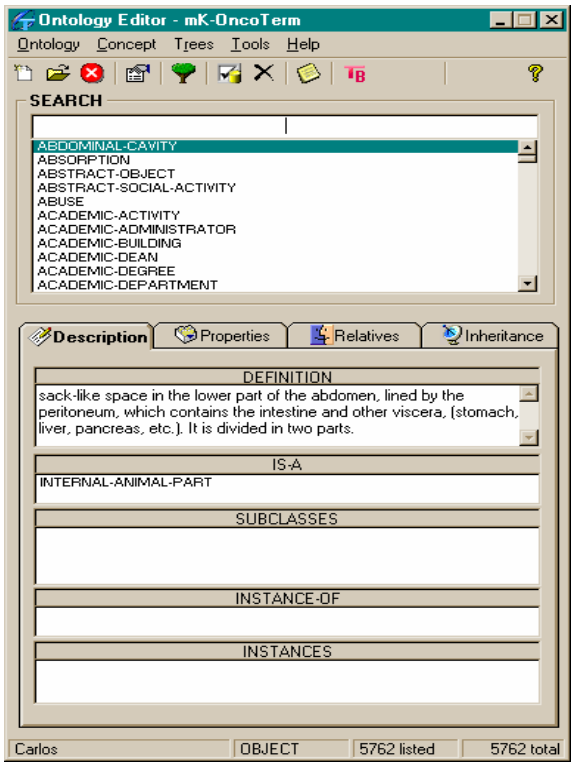
ejemplo, el momento en que se administra la radiación en combinación con otros tratamientos (*preoperative radiation therapy*, *postoperative radiation therapy*), o su función en una fase del desarrollo de la enfermedad (*curative radiation therapy*, *palliative radiation therapy*).

En el proyecto de investigación ONCOTERM se utiliza tanto la información extraída de diccionarios como la de textos especializados para la elaboración de una base de datos terminológica, en la que los conceptos están vinculados a una ontología. Dicha estructura conceptual subyace a todos los demás componentes y constituye el vínculo entre términos en diferentes lenguas.

Los datos quedan representados y almacenados en un gestor de conocimiento terminológico llamado Ontoterm (www.ontoterm.com), aplicación informática que ha desarrollado e implementado uno de los miembros de nuestro grupo, el Dr. Moreno Ortiz (Moreno Ortiz 2000ab; Moreno Ortiz y Pérez Hernández 2000). La necesidad de elaborar un nuevo recurso para el proyecto fue el resultado de las deficiencias percibidas en las bases de datos existentes, que, aun careciendo de un sistema conceptual adecuado, gozan de una gran difusión entre los usuarios de herramientas de gestión terminológica. La estructuración conceptual utilizada en esta aplicación informática depende de un recurso de representación de conocimiento ya existente, en concreto, la ontología Mikrokosmos (Mahesh y Nirenburg 1995, Viegas et al. 1999), desarrollada en el *Computing Research Laboratory* de la *New Mexico State University*. Estamos en proceso de ampliar los alrededor de 5000 conceptos iniciales para incluir y organizar conocimiento especializado sobre oncología. En la actualidad nuestra ontología cuenta con 5833 conceptos, de los que 397 son relaciones que, como hemos dicho, son las que nos van a permitir lograr un modelado sólido.

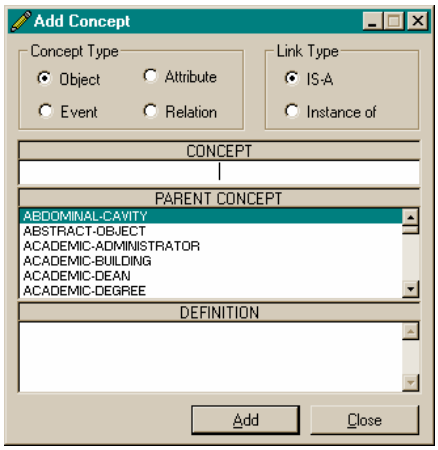
La arquitectura de Ontoterm tiene dos módulos principales: un editor de ontologías y un editor de la base de datos teminológica. En el editor de ontologías, se construye la estructura conceptual:

(4) Editor de ontologías



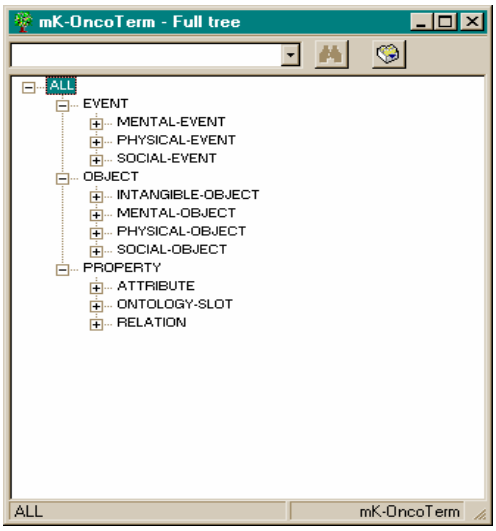
El editor de ontologías muestra la lista de conceptos en orden alfabético. Cada uno aparece con una descripción, propiedades, relaciones con otros conceptos y herencia. Este módulo tiene dos submódulos. El primer submódulo (5) se utiliza para añadir nuevos conceptos a la ontología. Para ello es necesario indicar el concepto superordinado (*parent concept*), con lo que queda especificado su lugar dentro de la ontología.:

(5) Submódulo para la integración de conceptos



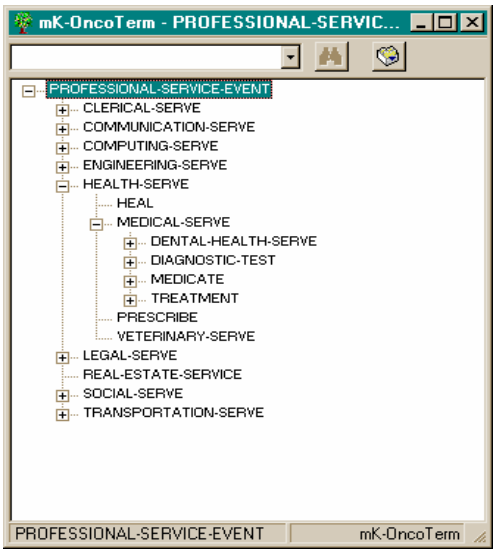
La jerarquía conceptual elaborada puede visualizarse en Ontoterm mediante la activación del segundo submódulo en el que todos los conceptos de la ontología aparecen en un árbol conceptual. A continuación, mostramos los conceptos más genéricos de la ontología de Mikrokosmos, en los que encajan todos los demás. Y así, cualquier concepto ha de clasificarse como EVENT, OBJECT o PROPERTY[4].

(6) Submódulo para la visualización de conceptos: conceptos genéricos



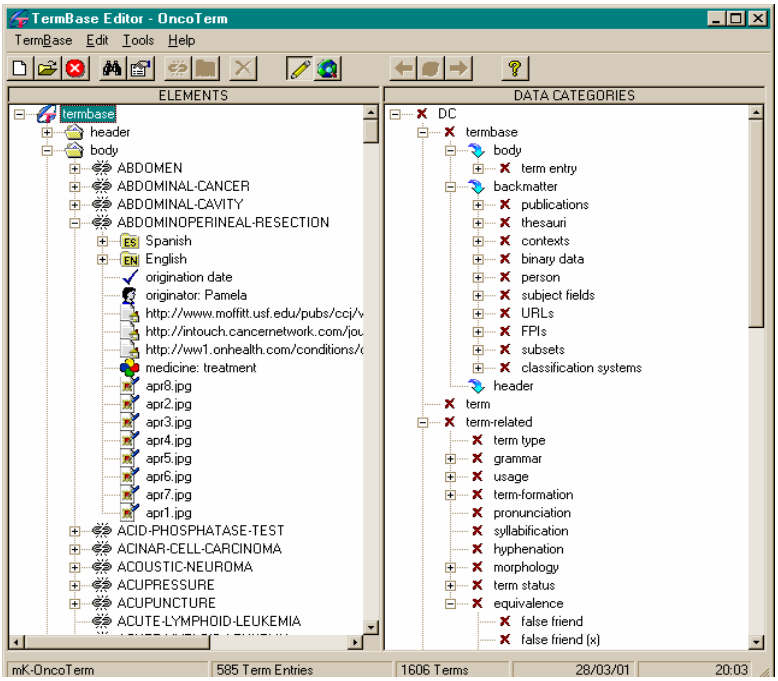
Con este programa, los conceptos especializados quedan integrados en los niveles más específicos de la ontología. Por ejemplo, conceptos muy importantes dentro del dominio ONCOLOGY, como TREATMENT o DIAGNOSTIC-TEST, quedan ubicados como hijos de MEDICAL-SERVE.

(7) Submódulo para la visualización de conceptos: HEALTH-SERVE



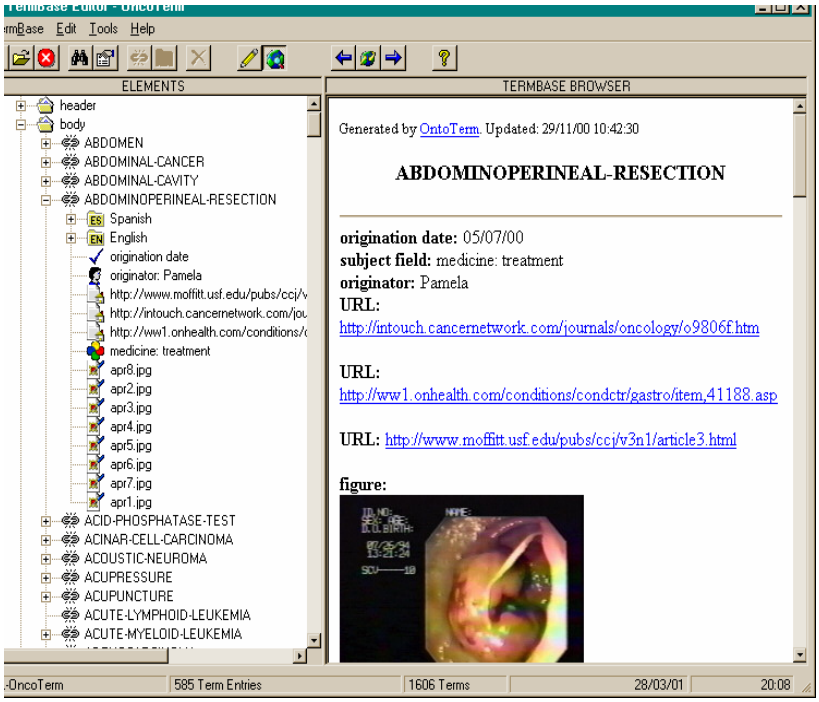
El segundo módulo de Ontoterm es el *editor de la base de datos terminológica*, donde se elaboran las entradas terminográficas vinculadas a los conceptos activados en la ontología. En este módulo, que complementa al editor de ontologías, se describe el término o términos que representan cada concepto mediante las categorías de datos de la norma ISO 12620.

(8) Editor de la base de datos terminológica



La información contenida en la base de datos terminológica también puede representarse más gráficamente en formato *HTML*, donde figuran todos los datos pertinentes al concepto en cuestión.

(9) Entrada terminológica en HTML



No obstante, en este artículo nuestro propósito no es tanto mostrar la aplicación informática utilizada, sino describir el tipo de análisis que seguimos en la extracción de datos lingüísticos y en la estructuración del conocimiento especializado. Dicha información proviene de corpora monolingües y bilingües.

2.2. Descripción del corpus

En ONCOTERM se combina un corpus paralelo y un corpus comparable[5] en inglés y español sobre oncología de aproximadamente 32 millones de palabras, de las cuales la mayor parte está en inglés (28.771.714). En la compilación del corpus se han utilizado: a) textos extraídos de Internet; b) enciclopedias, manuales y publicaciones médicas en CD-ROM y c) textos escaneados por su riqueza de vocabulario definicional (manuales para médicos y estudiantes de medicina) y por su relevancia entre especialistas, según MEDLINE, la base de datos bibliográfica más consultada por estos.

Teniendo en cuenta que los usuarios potenciales de nuestro proyecto son traductores e intérpretes, redactores técnicos, creadores de aplicaciones para el procesamiento del lenguaje natural, profesionales de la salud, pacientes de cáncer y público general interesado en la oncología, en la recopilación del corpus hemos pretendido reflejar las siguientes situaciones comunicativas:

(10) Situaciones comunicativas reflejadas en el corpus

Para plasmar dichas situaciones, se han consultado, por tanto, las páginas web de las principales organizaciones internacionales contra el cáncer[6], que ofrecen, tanto a profesionales de la salud como al público general, folletos informativos, resúmenes y publicaciones especializadas. También hemos seleccionado artículos experimentales y resúmenes provenientes de publicaciones médicas y oncológicas[7], y fragmentos relevantes de manuales destinados a oncólogos y estudiantes de medicina de nivel avanzado[8]. Por último, para reflejar la comunicación entre el especialista y un público lego, hemos recopilado no sólo monográficos que aparecen en enciclopedias y publicaciones de divulgación semiespecializada y general[9], sino también folletos de salud pública para pacientes y familiares disponibles en los sitios web antes mencionados (Tercedor 1999, Pérez Hernández 2000 y López Rodríguez 2001).

Entre los criterios de selección del corpus, Pérez Hernández (2000: 179-184), uno de los miembros de nuestro equipo, ha destacado los siguientes: cantidad, calidad, simplicidad, documentación, pertenencia al dominio de especialidad, fecha de producción y condición lingüística del texto, factualidad, tipo textual, nivel de tecnicidad y receptores del texto.

2.3. Aplicaciones para extraer datos de los corpora

Nuestra explotación de la lingüística de corpus para extraer términos e información especializada se fundamenta en el análisis de listas de frecuencia, líneas de concordancia y tablas de colocaciones.

Las *listas de frecuencia* no sólo proporcionan la frecuencia absoluta (*Frec*) y relativa (*%*) del término estudiado. Si éstas se lematizan, es decir, si las variantes morfológicas de un mismo lexema se agrupan, y se asocian estos lemas con otros de igual significado, es posible percibir las áreas conceptuales más activadas en un texto o

en un corpus (López Rodríguez 2001). En la lista de frecuencia (11), extraída con el programa *Wordsmith Tools*[10] a partir de nuestro corpus, destacan palabras de la lengua general que hacen referencia principalmente a los conceptos PATIENT, CELL, TREATMENT, DISEASE, RESEARCH, CAUSE y EFFECT. De ahí es posible inferir que estas son algunas de las áreas conceptuales en torno a las cuales se articula todo el conocimiento especializado del subdominio de la oncología, algo que también contribuye al modelado conceptual de la ontología que estamos creando.

(11) Lista de frecuencia correspondiente al corpus en inglés

N	PALABRA	FREC.	%	LEMAS
1	PATIENTS	175.546	0,61	patient(23100),patient's(2115)
2	CELLS	144.895	0,50	cell(65679),cell's(75)
3	JOURNAL	97.597	0,34	
4	ARTICLE	96.714	0,34	
5	STUDY	81.291	0,28	
6	CANCER	81.160	0,28	cancer's (12)
7	TREATMENT	62.967	0,22	treatments (4012)
8	RESULTS	61.455	0,21	result(7130)
9	USING	61.122	0,21	used(28489)
10	GROUP	54.309	0,19	groups (17340)
11	EFFECTS	52.790	0,18	effect(25848)
12	TUMOR	50.779	0,18	tumors(17569),tumour(4870), tumor's(41), tumour's(4)
13	DISEASE	49.773	0,17	diseases(6231),disease's(10)
14	PROTEIN	49.555	0,17	proteins(13187)
15	SIGNIFICANTLY	47.642	0,17	significant(23819)
16	INCREASED	44.978	0,16	increased (17955) , increases (215), increasing (64)
17	CLINICAL	42.253	0,15	
18	CASE	40.768	0,14	cases(26873)
19	YEARS	40.575	0,14	year(14623)
20	THERAPY	39.985	0,14	therapies(1528),therapeutic(5451),therapeutical(145), therapeutics(238), therapic(2)
21	SHOWED	34.180	0,12	shown (9456), show (3420)
22	ANALYSIS	32.539	0,11	analyses(3640),analysed(2036)
23	HIGH	32.051	0,11	
24	ACTIVITY	30.957	0,11	
25	HUMAN	30.603	0,11	
26	EXPRESSION	30.151	0,10	
27	TYPE	30.033	0,10	types(6896)
28	TIME	29.337	0,10	times(6594)
29	GENES	28.069	0,10	
30	DOSE	28.758	0,10	doses (6424)
31	LEVELS	27.813	0,10	
32	COMPARED	27.631	0,10	compare(3036),comparing(1749),compares(306)
33	INDUCED	27.420	0,10	induce(3069)
34	ASSOCIATED	26.053	0,09	
35	FOUND	25.900	0,09	find (9867)
36	RISK	25.740	0,09	risks(1705)
37	AGE	25.185	0,09	
38	TREATED	24.345	0,08	treat(2100),treating(1438)
39	RESPONSE	24.249	0,08	
40	CONTROL	23.878	0,08	
41	DATA	22.487	0,08	
42	NORMAL	22.310	0,08	
43	CHEMOTHERAPY	21.509	0,07	chemotherapeutic(808),chemotherapies(34),chemotherapy's(3)
44	MG	21.445	0,07	
45	SPECIFIC	21.370	0,07	

Las palabras que encabezan las listas de frecuencia de un texto o corpus también constituyen las representaciones lingüísticas de los conceptos que serían los más genéricos de nuestro dominio médico, algo que tiene consecuencias terminográficas. Por ejemplo, *treatment* o *therapy* son la parte nuclear o *definiens* de las definiciones de *chemotherapy, radiation therapy* o *radiotherapy*. Aunque distintos diccionarios médicos discrepan en cuanto al tipo y cantidad de información especificada para describir el mismo concepto, coinciden en ubicar el concepto de RADIATION THERAPY dentro de la categoría TREATMENT, como se muestra a continuación:

(12) Definiciones de *radiation therapy*

RADIATION THERAPY/RADIOTHERAPY	
HarperCollins Medical Dictionary	the **treatment** of disease by any radioactive substance or radiant energy.
The Cancer Dictionary	the use of high-energy penetrating rays or subatomic particles to **treat** or control disease.
Stedman's Concise Medical Dictionary	**treatment** with x-rays or radionuclides.
On-line Medical Dictionary (www.graylab.ac.uk/omd)	**treatment** with high energy radiation from X-rays or other sources of radiation

En consecuencia, a partir de los lemas más frecuentes de un corpus es posible identificar las categorías conceptuales sobre las que se fundamenta la definición de los términos del texto. Asimismo, las líneas de frecuencia también sirven para identificar patrones que servirán de nodos en líneas de concordancia de las que se extraerá información para las definiciones de los términos (Pearson 1998, López Rodríguez 2001).

(13) Líneas de concordancia en torno a *called*

```
1     e product of the body's cells. Tubes called bronchi make up the inside o
2     ancer: squamous cell carcinoma (also called epidermoid carcinoma), adeno
3     oat and into the bronchi. This test, called bronchoscopy, is usually don
4     e chest between two ribs. This test, called thoracoscopy, is usually don
5      also may be removed in an operation called a pneumonectomy. Sometimes p
6      the vein or muscle. Chemotherapy is called a systemic treatment because
7     One new type of radiation therapy is called radiosurgery. In radiosurger
8     out only a small part of the lung is called a wedge resection. When a wh
9      the vein or muscle. Chemotherapy is called a systemic treatment because
11    . Like surgery, radiation therapy is called local treatment because it w
12     lung is taken out, the operation is called a lobectomy. When one whole
13    n one whole lung is taken out, it is called a pneumonectomy.   Radiation
```

(López Rodríguez 2001: 513)

En (13) puede observarse que el radio colocacional de **called** aporta información sobre relaciones jerárquicas entre conceptos (BRONCHOSCOPY es un DIAGNOSTIC_TEST).

Asimismo la identificación de estos patrones nos ayuda en la extracción de términos. En este caso, el término es siempre el que aparece a la derecha de **called** (*pneumonectomy, lobectomy, radiosurgery*, etc.).

Como mencionamos previamente en (3), las *líneas de concordancia* ilustran el contexto de la unidad de significación especializada objeto de estudio y ayudan en la modelación del conocimiento. Estas pueden complementarse con *tablas de colocaciones* en las que se especifican lexemas con los que coocurren. A partir de estas se pueden detectar, por una parte, términos relacionados y, por otra, las unidades fraseológicas y los patrones sintácticos en los que encajan los términos (Tercedor 1999: 218-221), como se percibe en el ejemplo.

(14) Patrones sintácticos (español e inglés) de algunas colocaciones frecuentes del área de la oncología

N + Adj →N + N / Adj + N
progresión tumoral → tumor progression
agente cancerígeno → cancer agents
iniciador tumoral → tumor initiator
promotor tumoral → tumor promoter
suicidio celular → cell suicide / cellular suicide
Braquioterapia endobronquial → endobronchial brachytherapy
Anticuerpos monoclonales → monoclonal antibodies

3. Resultados: la organización conceptual en oncología

La organización de conceptos es una actividad que debe ser llevada a cabo según una base teórica sólida. Una manera de estructurar conocimiento consiste en la elaboración de una lista intuitiva de dominios y subdominios conceptuales *ad hoc*, a la que, posteriormente, lexicógrafos o terminógrafos asignan las unidades léxicas o terminológicas que piensan que son las más adecuadas. Si bien este tipo de método *top-down* es muy frecuente en la elaboración de tesauros y bases de datos léxicas, no necesariamente asegura una organización fiable.

Es preferible que la gestión de terminología no sea un proceso cien por cien *top-down* o *bottom-up*, sino una mezcla de los dos. Esto se debe al hecho de que un dominio conceptual no puede representarse como un sistema totalmente abierto o cerrado. El conocimiento especializado es una subdivisión de nuestro sistema de conocimiento general. Por consiguiente, se compone de esquemas cognitivos flexibles que permiten la manipulación de ideas y la construcción de hipótesis. Sin embargo, tales esquemas

necesariamente conllevan restricciones, lo que implica un sistema de atributos, entidades y relaciones bien definido. En terminología esto supone la especificación de un metalenguaje que consiste en una estructura conceptual con relaciones preestablecidas entre conceptos, reflejadas en la definición de cada unidad terminológica. En este sentido, las definiciones terminográficas se conciben como el puente entre conceptos y términos.

La aplicación de teorías lingüísticas a la terminología no siempre ha tenido buenos resultados, ya que cualquier teoría válida para la terminología debe tener un importante componente semántico, capaz de dar cuenta no sólo del significado de unidades léxicas, sino también de sus relaciones con otras unidades con significado similar. Esto último es fundamental en terminología, donde la representación de relaciones conceptuales es esencial.

Por consiguiente, para la codificación de conocimiento especializado y el análisis de los datos del corpus, hemos utilizado el Modelo Lexemático Funcional (MLF) de Martín Mingorance (1984, 1989, 1995; Faber y Mairal 1999), teoría de base léxica que facilita la representación de relaciones conceptuales y colocacionales en el lenguaje general y especializado.

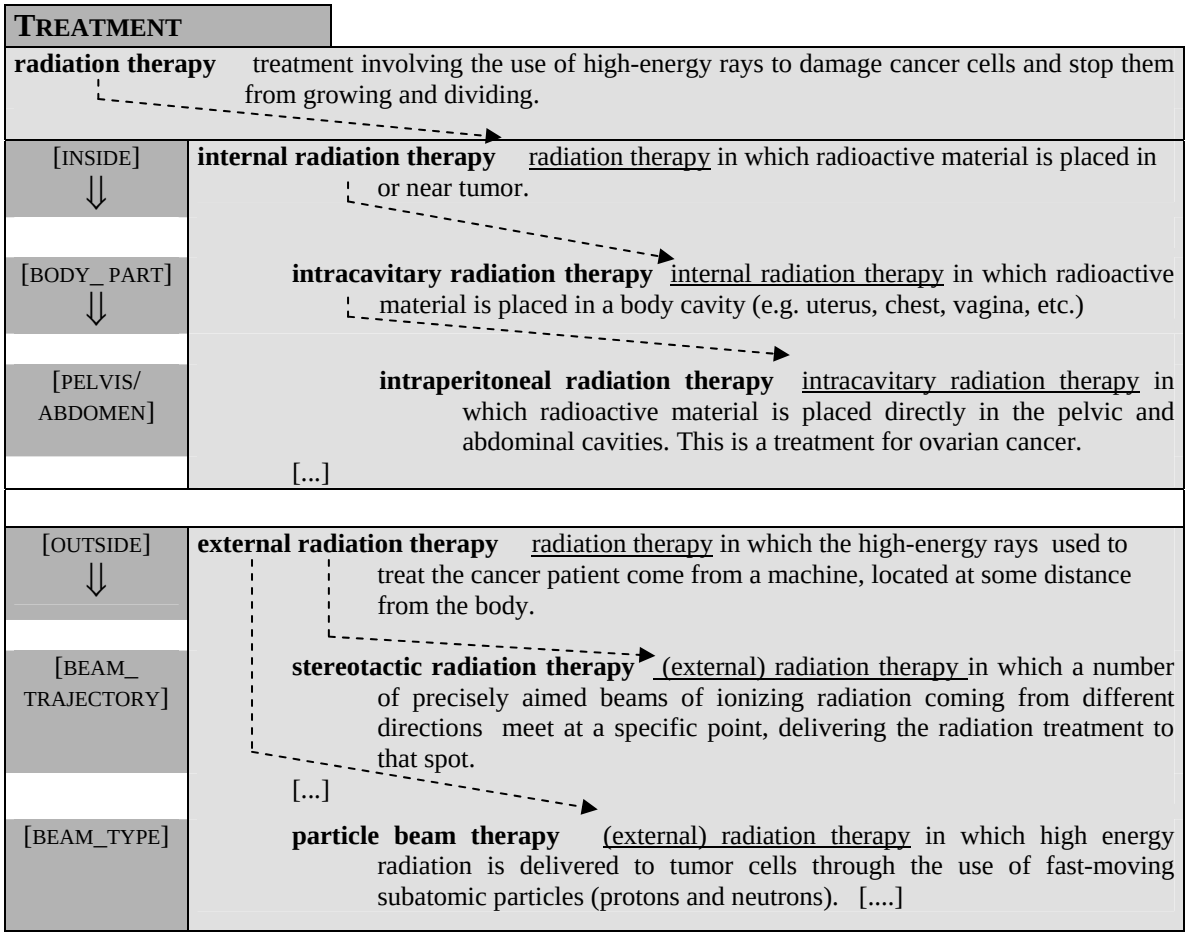
La organización léxica que el MLF propone para el lexicón se basa parcialmente en la distinción entre relaciones sintagmáticas y paradigmáticas, o los principios complementarios de combinación y selección (Saussure [1916] 1990; Lyons 1977: 241). Esta distinción es pertinente porque se encuentra en nuestra organización conceptual, previa e independientemente del sistema lingüístico (Nelson 1985: 179).

El eje paradigmático del lexicón del MLF codifica la configuración de conceptos en el eje de selección, organizándolos onomasiológicamente en una jerarquía de dominios y subdominios. Asimismo, es un factor determinante en el eje sintagmático, en el que se codifica su potencial combinatorio. La convergencia de estos dos ejes es la base de la estructura conceptual propuesta.

Este tipo de enfoque relacional pone el énfasis en el significado, y más concretamente, en las áreas conceptuales. Se basa en la premisa de que existen una serie de propiedades compartidas por todos los miembros de un dominio conceptual, y otras que los diferencian entre sí. Este modelo concibe la representación de la memoria semántica como una compleja red en la que cada nodo es un concepto, y los conceptos están interconectados por diferentes tipos de relaciones (Iris, Litowitz y Evens 1988: 263).

El MLF ofrece una metodología lingüística para la organización de conceptos, mediante la información extraída de definiciones lexicográficas / terminográficas estructuradas sistemáticamente. En concreto, la estructuración de RADIATION THERAPY, como subdominio de TREATMENT, tendría el siguiente diseño, explícito en las definiciones de los términos:

(15) Jerarquía conceptual de RADIATION THERAPY reflejada en las definiciones de los términos



De este modo, la estructuración conceptual queda reflejada en definiciones terminográficas consensuadas, que han sido reelaboradas, utilizando los componentes de significado encontrados tanto en el corpus, como en diccionarios médicos de reconocido prestigio. En este sentido, en ONCOTERM compartimos la opinión de Bejoint (1997: 19-20), quien señala que en terminología nunca se ha dado la debida importancia a las definiciones, que no deben considerarse como información dada, sino como construcciones.

3.1. Sustantivos terminológicos

Al analizar nuestro corpus sobre oncología, hemos identificado ocho grandes categorías conceptuales, muchas de ellas extrapolables a otros dominios médicos (Faber 1999: 99; Faber y Mairal 1999):

(16) Categorías conceptuales en el EVENTO MÉDICO

- ❑ BODY_PART
- ❑ TREATMENT
- ❑ DRUG
- ❑ INSTRUMENT
- ❑ TUMOR
- ❑ SPECIALIST
- ❑ HOSPITAL
- ❑ DIAGNOSTIC_PROCEDURE

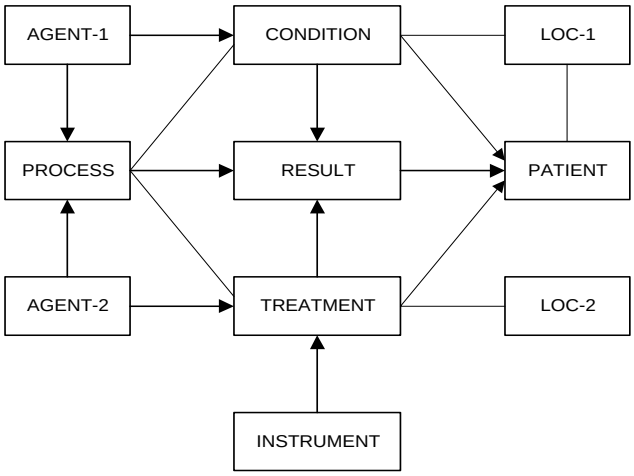
Hemos integrado estas categorías en la Ontología Mikrokosmos, extendiéndola para incluir conocimiento especializado. En un principio se pensó que *risk factor, symptom* y *side effect* podrían ser categorías conceptuales dentro de la ontología. No obstante, al constatar que ASBESTOS, SMOKING y SUNLIGHT (*factores de riesgo*) o COUGH y FATIGUE (*síntomas*) ya formaban parte de ella, se decidió representar estos conceptos como relaciones, por ejemplo, SYMPTOM-OF y su inversa HAS-SYMPTOM:

(17) Ejemplos de relaciones conceptuales

	RELACIÓN CONCEPTUAL ⇓	
COUGH	**SYMPTOM-OF**	LUNG_CANCER
LUNG_CANCER	**HAS-SYMPTOM**	COUGH
SMOKING	**RISK-FACTOR-OF**	LUNG_CANCER
LUNG_CANCER	**HAS-RISK-FACTOR**	SMOKING

Estas categorías se integran en un diagrama que Faber (op. cit.) denomina EVENTO MÉDICO.

(18) EVENTO MÉDICO



Las denominaciones lingüísticas de estas áreas conceptuales pertenecen a la lengua general, aunque sus niveles más específicos son propios del campo del saber en cuestión, algo que refuerza la idea de que en los campos de especialidad no existen claras fronteras entre los lexemas de la lengua general y las unidades terminológicas (Cabré 1999).

La estructura interna de cada categoría conceptual está representada por un conjunto de tipos de información. Cada categoría contiene información sobre las características de un concepto (intensión) y también pone de manifiesto una descripción de entidades del mundo real que pueden pertenecer a esta categoría (extensión). Por lo tanto, cada concepto posee un esquema básico que sirve de modelo de la categoría conceptual en cuestión. El esquema básico para la categoría conceptual TREATMENT sería el siguiente:

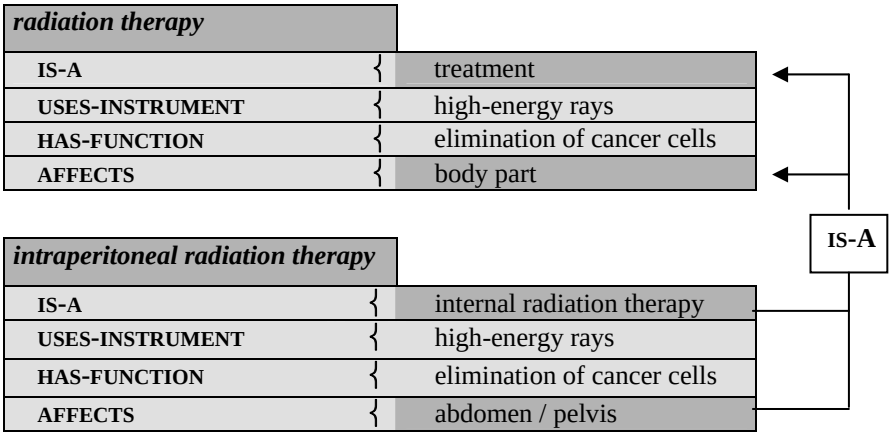
(19) Esquema categorial de TREATMENT

CONCEPTUAL CATEGORY	CONCEPTUAL RELATION
TREATMENT	**IS-A**
	USES-INSTRUMENT
	HAS-FUNCTION
	HAS-LOCATION

Si se aplica el esquema básico de TREATMENT a los términos que pertenecen al subdominio RADIATION THERAPY, como *intraperitoneal radiation therapy*, se percibe que hay una herencia de todos los valores generados a partir de las relaciones que configuran el esquema. Más concretamente, se puede observar que *intraperitoneal radiation therapy* IS-A *internal radiation therapy* IS-A *radiation therapy* IS-A *treatment*. Igualmente,

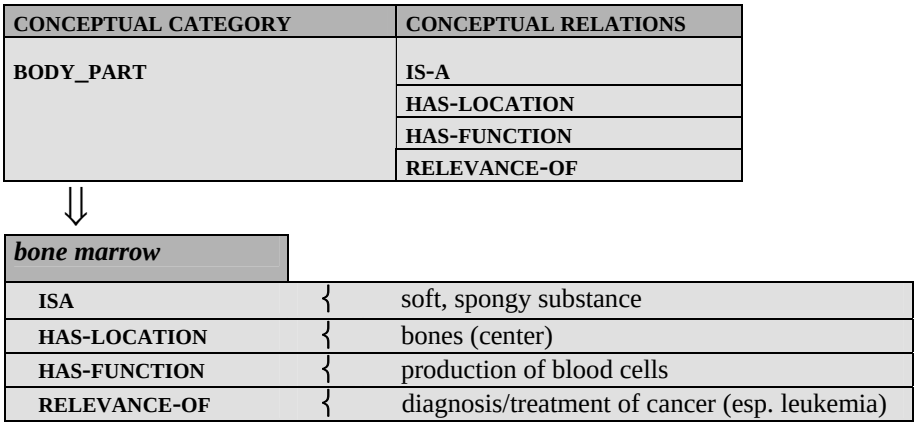
intraperitoneal radiation therapy AFFECTS *abdomen/pelvis,* con lo que se hereda el esquema *radiation therapy* AFFECTS *body part.*

(20) Aplicación del esquema categorial de TREATMENT a *radiation therapy* y *intraperitoneal radiation therapy*



Partimos pues de la hipótesis de que cada categoría conceptual lleva su esquema prototípico de relaciones conceptuales. Esta información conceptual básica puede utilizarse para "formatear" la definición de términos, aumentando así su comprensión y expresando las conexiones con otros conceptos explícitos. Veamos ahora el esquema para la categoría BODY_PART y su aplicación al término *bone marrow*.

(21) Esquema categorial de BODY_PART y su aplicación a *bone marrow*



Puesto que cada categoría tiene su propio patrón conceptual, las definiciones de los términos para diferentes lenguas pueden formularse partiendo de la misma

configuración de información subyacente. De esta manera, las definiciones ganan en homogeneidad y coherencia, al mismo tiempo que plasman las relaciones conceptuales.

(22) Definición de *bone marrow / médula ósea*

bone marrow
soft and spongy substance **[ISA]** in the centre of bones **[HAS-LOCATION]** which is responsible for the production of blood cells, and particularly of red cells and platelets **[HAS-FUNCTION]**. In some varieties of cancer, especially leukaemia, the bone marrow produces abnormal blood cells. It is, therefore, an important element for the diagnosis and treatment of cancer **[RELEVANCE-OF]**.
médula ósea
sustancia blanda y esponjosa **[ISA]** situada en el centro de los huesos **[HAS-LOCATION]**, cuya función es la producción de células sanguíneas, en particular de glóbulos rojos y plaquetas **[HAS-FUNCTION]**. En algunas variedades de cáncer, especialmente en la leucemia, la médula ósea produce células sanguíneas anormales, por lo que resulta un elemento importante en el diagnóstico y tratamiento del cáncer **[RELEVANCE-OF]**.

El análisis de las líneas de concordancia en torno a *bone marrow* nos ayudará a ubicar estas categorías conceptuales en la ontología, puesto que ponen de manifiesto que el objeto BODY_PART y el evento TREATMENT están vinculados mediante la relación RELEVANCE-OF. De las líneas de concordancia inferimos que la médula ósea (*bone marrow*) es relevante para el tratamiento de la leucemia, ya que puede observarse su coocurrencia con *transplantation* y otros términos relacionados con el tratamiento. También extraemos las denominaciones de los conceptos correspondientes a los posibles subtipos de este tratamiento (*allogeneic bone marrow transplantion*, *autologous bone marrow transplantation*, *syngeneic bone marrow transplantation*, etc.).

(23) Líneas de concordancias a partir de *bone marrow*

BONE MARROW

```
N       transplantation: A procedure to replace bone marrow destroyed by treatment
1       39% +/- 7% at 5 years) after allogeneic bone marrow transplantation.
2       MC, et al.: BAVC regimen and autologous bone marrow transplantation in
3       addition, total-body irradiation before bone marrow transplant increases
4        Allogeneic versus autologous    purged bone marrow transplantation for
5       has led some investigators to recommend bone marrow or peripheral stem
6       transplantation: A procedure to replace bone marrow destroyed by treatment
7       us 32%).[Level of evidence: 1A] Salvage bone marrow transplantation was
8       ot al.: Should HLA-identical sibling    bone marrow transplants for
9       ion from an   identical twin (syngeneic bone marrow transplantation). This
```

3.2. Los verbos en el discurso especializado

Cuando se habla de términos, normalmente se centra la atención en los sustantivos. Sin embargo, en la comprensión y estructuración del discurso especializado, al igual que en la lengua general, los verbos desempeñan un papel muy importante. De hecho, gran parte de nuestro conocimiento está constituido por EVENTOS y ESTADOS, muchos de ellos representados lingüísticamente por verbos.

En un proyecto de investigación anterior[11], elaboramos una red semántica representativa del significado conceptual subyacente en el lexicón verbal de la lengua general. Nuestro objetivo principal era investigar el potencial que ofrece la arquitectura léxica de una lengua para la representación del conocimiento. Para ello, analizamos y clasificamos onomasiológicamente alrededor de 8 000 verbos en inglés y español en campos léxicos jerarquizados según las premisas del Modelo Lexemático Funcional (Faber y Mairal 1998, 1999). Se estableció un sistema de definiciones basado en el *Decomposition Principle* de Mel'cuk (1988), que estipula que la definición de una unidad léxica **L** debe contener únicamente términos que sean más simples que **L.** Wierzbicka (1992: 11) escribe al respecto:

> Explicating involves reducing semantically complex words to semantically simple words, and hence the words used in an explication are not selected at random: there is a hierarchy among words, and a correct definition will reflect this hierarchy.

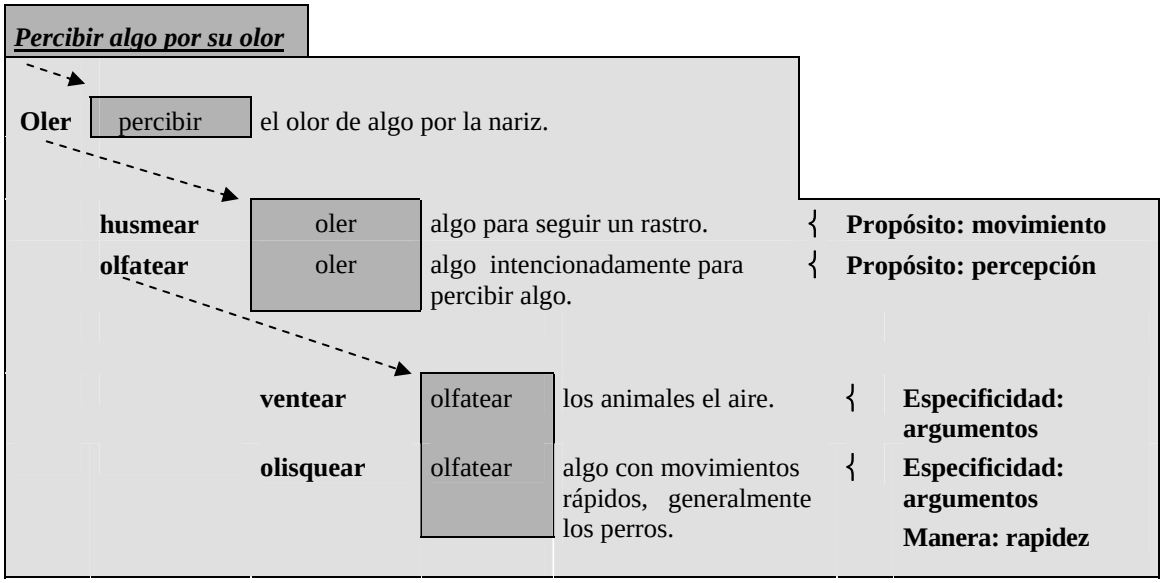
En nuestro diseño del lexicón, un dominio léxico es una jerarquía de lexemas que comparten, todos ellos, el mismo significado nuclear. Este núcleo de significado compartido marca el territorio semántico que abarca un determinado dominio o subdominio, por lo que se convierte en el factor que determina la pertenencia de un lexema a un área de significado conceptual.

Los lexemas especifican el significado nuclear del dominio al que pertenecen de diferentes formas, a través de sus *differentiae,* que a su vez representan parámetros semánticos, fieles reflejos de nuestra percepción. Las definiciones consensuadas de cada lexema se confeccionan mediante el análisis de corpus y la extracción de la información explícita en las entradas de varios diccionarios. (Faber y Mairal 1998: 237).

Cada dominio léxico posee uno o dos términos superordinados a partir de los cuales se definen los demás miembros del dominio. Los términos superordinados no coinciden con todo el contenido semántico de un dominio, sino que son los núcleos de contenido que se parametrizarán de diferentes maneras en los lexemas específicos de la

jerarquía. A continuación, presentamos un segmento del dominio PERCEPCIÓN. Especificamos los *differentiae* de los verbos subordinados a la derecha.

(24) Jerarquía de verbos de percepción olfativa



Igualmente, los patrones de complementación que toman los verbos pertenecientes al mismo dominio léxico son similares. Información sintagmática como el número y tipo de argumentos está semánticamente motivada, en el sentido de que las diferentes realizaciones sintácticas se conciben como proyecciones de los parámetros semánticos definitorios de cada dominio y subdominio. En (24) puede verse que los lexemas con un significado más específico tienen patrones de complementación más restringidos, sobre todo en cuanto a los rasgos semánticos de sus argumentos.

En el caso de la comunicación especializada, el verbo también estructura el discurso, determinando el número de argumentos en cada proposición, así como su naturaleza y función. Los argumentos en el discurso especializado suelen ser términos, y los verbos son determinantes en su configuración. En efecto, los predicados, y más concretamente, el significado que codifican hacen explícitos las relaciones conceptuales entre las entidades en un campo especializado. La focalización en una parte determinada de un dominio léxico de la lengua general resalta determinados argumentos que, como veremos, en el discurso especializado activan jerarquías enteras de conceptos.

Sin embargo, el uso de estos verbos difiere considerablemente con respecto a su uso en la lengua general. Es bien sabido que los términos (sustantivos) suelen ser monosémicos, pero la misma restricción de significado también alcanza a los verbos de

la lengua general utilizados en los discursos especializados. Dichos predicados también restringen su significado en cuanto a la naturaleza semántica de los argumentos que pueden aparecer en su cotexto. Por ejemplo, el verbo *respond* tiene los siguientes significados, que quedan reducidos en el discurso especializado sólo al último de ellos.

(25) Significados de *respond*

RESPOND (lengua general)	
Longman Dictionary of English Language and Culture	**1** to say or write (something) in reply
	2 to do something in answer.
	3 to get better as a result of a treatment.
RESPOND (discurso especializado)	
Líneas de concordancia	⇨ to get better as a result of a treatment.

Desarrollamos las líneas de concordancia que constatan fácilmente la restricción del significado que experimentan los verbos en el discurso especializado:

(26) Líneas de concordancia correspondientes a *respond*

RESPOND

1	r estrogen and progesterone receptors **respond** best to progestin therapy. Among
2	aid that stage I and stage II cancers **respond** equally well to radiation or sur
3	onal cell cancer of the urethra may **respond** favorably to the same chemothera
4	ith tumors confined to the cervix which **respond** incompletely to radiation ther
5	with MGUS or smoldering myeloma do not **respond** more frequently, achieve longe
6	trials are underway. For patients who **respond** to their initial therapy, the my
7	of evidence: 3iiiDi] For patients who **respond** to neoadjuvant chemotherapy, loc
8	w transplantation, some patients will **respond** to interferon alfa.6 Infusions
9	onding patients who relapse usually **respond** to retreatment with interferon a
10	after relapse. The primary group may **respond** to high-dose chemotherapy and au
11	atonin may affect the way tumor cells **respond** to chemotherapy and radiation th
12	ortunistic infection.14 Most patients **respond** to treatment by showing partial
13	nd secondary refractory patients who do **respond** to induction chemotherapy, but
14	esponse rate of 20% in those who do not **respond** to standard progesterone therapy
15	topenia or hemolytic anemia who fail to **respond** to alkylating agents and pre
16	resected or metastatic tumors failed to **respond** to chemotherapeutic agents f
17	r other mechanical problems expected to **respond** to antineoplastic therapy
18	Tamoxifen: Some patients (18%) will **respond** to tamoxifen (20 milligrams
19	bHCG and AFP. Certain of these tumors **respond** to platinum-based combination ch
20	astases from small cell carcinoma may **respond** to chemotherapy as readily as me
21	" tumor resection. Low-grade tumors may **respond** to various chemotherapeutic re
22	ow and serious infections that do not **respond** to antibiotics.20 Prophylactic o
23	on: Patients whose disease does not **respond** to combined radiation therapy an
24	effective.1,3,4 Patients who do not **respond** to a cisplatin-based combination
25	nodal relapse. Patients who do not **respond** to induction chemotherapy (about
26	more effective.1-3 Patients who do not **respond** to a cisplatin-based combination
27	indolent lymphoma: half of patients **respond** to a four-dose treatment program

28	TIC – to treat painful conditions that **respond** to nerve blocks (e.g., celia
29	y is rare.4 All patients who do not **respond** to standard therapy are candidat
30	emotherapy at the time of CNS relapse **respond** to second-line chemotherapy.13
31	the patients treated. Patients who **respond** usually demonstrate improvement
32	with metastatic disease at diagnosis **respond** well to the therapy given to pat
33	is continued; if the tumor did not **respond**, an alternative regimen is used.
34	th newly diagnosed medulloblastoma will **respond**, at least partially, to chemot
35	ge with newly diagnosed ependymoma will **respond**, at least partially, to chemothe
36	dermal tumors and pineoblastomas will **respond**, at least partially, to chemothe
37	and vincristine.1 If the tumor fails to **respond**, it may be a benign lesion.
38	and reticulocyte counts usually do not **respond**. The effect of GM-CSF treatment
39	small cell lung cancers that have not **responded** to other anticancer drugs.
40	vement. The disease in these sites also **responded** to vinblastine. No studies
41	nous cisplatin-based chemotherapy and **responded** to second-line intraperitoneal
42	ds in unresectable tumors that have not **responded** to radiation therapy.
43	-refractory ovarian cancer who have not **responded** to paclitaxel (Taxol) ma
44	sponded to initial therapy and who have **responded** to conventional therapy for
45	f initial diagnosis. For patients who **responded** to first-line chemotherapy but
46	supportive care. In those patients who **responded** to low-dose cytarabine, respon
47	th metastatic breast cancer who are **responding** to conventional induction
48	low-grade lymphomas, but this lymphoma **responds** less well to chemotherapy tha

Las concordancias en (26) muestran que *respond* sólo aparece con el significado de *to get better as a result of a treatment.* En estos textos, *respond* es un predicado de dos argumentos, cada uno de los cuales tiene un rol semántico característico y pertenece a una categoría conceptual específica:

(27)

Primer argumento	Predicado	Segundo argumento
(CONDITION)$_{\text{AFFECTED-BY}}$	**RESPOND**	(TREATMENT)$_{\text{AFFECTS}}$

Además, si analizamos los datos en (26), las unidades que rellenan las casillas de los argumentos son términos que forman jerarquías. Dichas jerarquías activan determinados sectores de una ontología de conceptos propia del dominio oncológico. Por tanto, la información extraída de nuestro corpus proporciona la base para la estructura conceptual de este dominio. En (28) se han reagrupado los datos de las concordancias para hacer explícita una sección de la estructura de la categoría conceptual TREATMENT. De aquí saldrán eventos, objetos y relaciones que quedarán plasmados en el editor de ontologías.

(28) Conceptos activados por el verbo *respond**: estructuración de la categoría TREATMENT

Primer argumento	Predicado	Segundo argumento
DISEASE ↳ painful condition **CANCER** ↳ small cell lung cancer ↳ metastatic breast cancer **TUMOR** ↳ tumor cells ↳ metastatic tumor ↳ unresectable tumor ↳ small cell carcinoma ↳ myeloma ↳ medulloblastoma ↳ ependymoma ↳ pinelblastoma ↳ lymphoma **PATIENT** ↳ cancer patient	**RESPOND**	**TREATMENT** ↳ retreatment ↳ second treatment ↳ four-dose treatment ↳ GM-CSF treatment **THERAPY** ↳ standard therapy ↳ initial therapy ↳ ***chemotherapy*** ↳ neoadjuvant chemotherapy ↳ high-dose chemotherapy ↳ second-line chemotherapy ↳ induction chemotherapy ↳ anticancer drug ↳ chemotherapeutic agent ↳ alkylating agent ↳ platinum-based combination ↳ cisplatin-based combination ↳ Cyclophosfamide ***radiation therapy*** ↳ combined radiation therapy ***biological therapy*** ↳ Interferon ↳ Interferon alfa 6

Este tipo de información valida el evento médico (véase apartado 3) y constituye una fuente muy valiosa para afinar el tipo de relaciones entre los participantes del mismo, de forma que se pueda extraer un macroesquema del dominio de la oncología. Al poner el énfasis en la especificación basada en corpus de relaciones conceptuales, establecemos una base empírica para la ontología y la base de datos terminológica que estamos desarrollando en el proyecto ONCOTERM.

4. Conclusiones

En este artículo hemos descrito de manera sucinta el proyecto de investigación ONCOTERM y la metodología utilizada para la configuración de categorías conceptuales y la

estructuración de conocimiento especializado dentro del dominio de la Oncología. Para estructurar este dominio y llegar a un inventario de categorías, nos hemos basado en las premisas del Modelo Lexemático Funcional y la lingüística de corpus.

Al igual que en el léxico primario, en el léxico especializado debe utilizarse no sólo la información de definiciones lexicográficas sino también la extraída de corpora. Es decir, estos dos métodos son complementarios para la construcción de jerarquías conceptuales de términos y constituyen una herramienta útil para una representación verdaderamente multidimensional. El análisis de concordancias y frecuencias derivadas de nuestro corpus de textos médicos sobre oncología, complementado con la información extraída de diccionarios médicos de reconocido prestigio, es la base sobre la que hemos llegado a los siguientes resultados:

- un inventario de categorías conceptuales propias de la oncología
- la integración de dichas categorías y los conceptos que las conforman en la ontología de Mikrokosmos
- esquemas de la estructura interna de cada categoría y su aplicación a conceptos más específicos, así como a la confección de definiciones terminográficas concisas y sistemáticas que reflejen la estructura conceptual del dominio de la oncología
- una metodología para el estudio del comportamiento de lexemas verbales en el discurso especializado médico.

En definitiva, la investigación terminográfica se facilita enormemente con la utilización de técnicas propias de la lingüística de corpus en el estudio y análisis de contextos. No obstante, sólo cuando estas técnicas se combinan con los principios de un marco teórico sólido y con la consulta a especialistas, es posible llegar a resultados fiables sobre la configuración del conocimiento especializado.

Notas

[1] Este trabajo se ha elaborado dentro del marco del proyecto de investigación *OncoTerm: Sistema bilingüe de información y recursos oncológicos*, financiado por el Ministerio de Educación y Cultura español (DGICYT, código PB98-1342).

[2] En un corpus sobre oncología, Tercedor (1999) extrae ejemplos de 50 funciones léxicas, que denomina funciones terminológicas.

[3] La ontología que nos ayuda a gestionar los conceptos, *Mikrocosmos* (véase 2.1.), define el primitivo semántico EVENT como "any activity, action, happening or situation".

[4] En *Mikrokosmos*, bajo OBJECT se agrupan "ontological concepts that are not actions, or properties. The static things that exist in the physical, mental, and social world", mientras que el concepto PROPERTY hace referencia a "the properties of entities or their states".

[5] Un *corpus paralelo* es aquel que presenta el mismo texto en más de una lengua, es decir, un texto y su traducción a una o más lenguas (McEnery 1996: 58). Por otra parte, un *corpus comparable* (Peters *et al.* 1996: 69) es un conjunto de textos en más de una lengua que, sin ser traducciones, por coincidir en el tema, motivación situacional y función comunicativa, proporciona una excelente base para la comparación de dos o más lenguas.

[6] Entre estos centros de información oncológica destacamos CancerNet, CancerBacup, Medscape, MedicineNet, Oncoweb, Virtual Hospital, Alcase, Atheneum y Diario Médico.

[7] En inglés, se incluyen textos publicados en *British Medical Journal, Lancet, New England Journal of Medicine, Cancer, CANCERLIT, C-A. A Cancer Journal for Clinicians*. Los textos en español provienen de *Medicina Clínica, Revista Clínica Española, Neoplasia, Revisiones en Cáncer, Revista Española de Anestesiología y Reanimación, Archivos Bronconeumológicos, Revista Española de Enfermedades Digestivas, Anales Otorrinolaringológicos Ibero-Americanos, Anales Españoles de Pediatría* y *Actas Urológicas españolas*.

[8] *Harrison's Principles of Internal Medicine, Cancer: Principles and Practice of Oncology, Medicina Interna de Farreras-Rozmán, Cancer. Principios y Práctica de Oncología* y *Oncología Médica-Guía de Oncología Médica.*

[9] Entre las enciclopedias médicas destinadas a un público no versado en medicina, podemos citar *The Merck Manual of Diagnosis and Therapy / Manual Merck en español* y *Mosby's Medical Encyclopedia for Health Consumers*. Hemos sacado también textos de la *Enciclopedia Microsoft Encarta 97*. Los monográficos provienen también de publicaciones de divulgación semiespecializada como *Scientific American/Investigación y Ciencia* y de divulgación general, tales como *Oncolink, Reuters Health,* la revista *TIME, QUO* y suplementos de periódicos (*Blanco y negro, El Mundo, El Semanal* o *Ideal*).

[10] Programa creado por Mike Scott [http://www.liv.ac.uk/~ms2928/homepage.html] de la *University of Liverpool* y distribuido por Oxford University Press.

[11] *Desarrollo de una lógica léxica para la traducción asistida por ordenador a partir de una base de datos léxica inglés-español-alemán*, financiado por el Ministerio de Educación y Cultura (PB 94-0437).

Bibliografía

Ahmad, K., A. Davies, H. Fulford y M. Rogers. 1994. "The elaboration of special language terms: the role of contextual examples, representative samples and normative requirements". *Euralex '92 Proceedings*. Tampere: Studia Translatologica, 139-150.

Amsler, R. A. 1980. *The structure of the Merriam-Webster Pocket Dictionary*. (Tech Rep. No. TR-164). Austin: University of Texas.

Béjoint, H. 1997. "Regards sur la définition en terminologie". *Cahiers de Lexicologie* 70 (1), 19-26.

Benson, M., E. Benson y R. Ilson. 1986. *The BBI Dictionary of English Word Combinations*. Amsterdam / Philadelphia: John Benjamins.

Bourigaut, D. 1994. *LEXTER, Un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir des textes*. Tesis doctoral. EHESS, París.

Bourigaut, D. y M. Slodzian. 1999. "Pour une terminologie textuelle". *Terminologies nouvelles* 19, 29-32.

Cabré Castellví, M.T. 1999. *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Daille, B. 1994. *Approche mixte pour l'extraction de terminologie: statistiques lexicales et filtres linguistiques*. Tesis doctoral. Université de Paris VII.

Faber, P. 1999. "Conceptual analysis and knowledge acquisition in scientific translation". *Terminologie et traduction* 1999(2), 97-123.

Faber, P. y R. Mairal Usón. 1998. "Dominios y esquemas de predicado: hacia una productividad léxica". En Wotjak (ed.), *Teoría del Campo y Semántica Léxica*. Frankfurt: Peter Lang.

Faber, P. y R. Mairal Usón. 1999. *Constructing a lexicon of English verbs*. Berlin: Mouton de Gruyter.

Faber, P. y M.I. Tercedor. 2001. "Codifying conceptual information in descriptive terminology management". *META* 46(1), 192-203.

Iris, M., B. Litowitz y M. Evens. 1988. "Problems of the part-whole relation". En Evens (ed.), *Relational Models of the Lexicon: representing knowledge in semantic networks*. Cambridge: Cambridge University Press, 261-287.

Jacquemin, C. 1997. "Recognition and acquisition. Two inter-related activities in corpus-based term extraction". *Terminology*, 4(2), 245-273.

Jacquemin, C. y J. Royauté. 1994. "Retrieving terms and their variants in a lexicalized unification-based framework". *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. Heidelberg: Springer-Verlag, 132-141.

Laporte, I. y M.C. L'Homme. 1997. "Recensemenet et consignation des combinaisons lexicales en langue de specialité: exemple d'application dans le domaine de la pharmacologie cardiovasculaire". *Terminologies nouvelles* 16, 95-101.

Lauriston, A. 1993. *Le repérage automatique des syntagmes terminologiques*, Tesina presentada en la Universidad de Québec.

L'Homme, M.C. 1996. "Definition of an evaluation grid for term-extraction software". *Terminology* 3(2), 291-312.

L'Homme, M.C., C. Bodson y R.S. Valente. 1999. "Recherche terminographique semi-automatisée en veille terminologique: experimentation dans le domaine médical". *Terminologies nouvelles* 20, 25-36.

López Rodríguez, C. 2001. *Tipología textual y cohesión en la traducción biomédica inglés-español: un estudio de corpus.* Tesis doctoral. Granada: Editorial Universidad de Granada.

Lyons, J. 1977. *Semantics*. 2 vols. Cambridge: Cambridge University Press

Mahesh, K. y S. Nirenburg. 1995. “A situated ontology for practical NLP”. En *Proceedings on basic ontological issues in knowledge sharing.* International Joint Conference on Artificial Intelligence (UCAI-1995), August 1995, Montreal, Canada.

Martín Mingorance, L. 1984. “Lexical fields and stepwise lexical decomposition in a contrastive English-Spanish verb valency dictionary”. En Hartmamann (ed.), *LEXeter 83: Proceedings of the International Conference on Lexicography*. Tübingen: Niemeyer, 226-236.

Martín Mingorance, L. 1989. “Functional Grammar and Lexematics”. En Tomaszczyk, J. y B. Lewandowska (eds.), *Meaning and Lexicography*. Amsterdam / Philadelphia: John Benjamins, 227-253.

Martín Mingorance, L. 1995. “Lexical logic and structural semantics: methodological underpinnings in the structuring of a lexical database for natural language processing. En Hoinkes (eds.), *Panorama der Lexikalischen Semantik.* Tubinga: Gunter Narr, 461-474.

McEnery, A. y A. Wilson. 1996. *Corpus Linguistics*. Edimburgo: Edinburgh University Press.

Meijs, W. and P. Vossen. (1992) “In so many words: knowledge as a lexical phenomenon”. En Pustejovsky, J. y S. Bergler (eds). *Lexical semantics and knowledge representation*. Berlin: Springer, 137-153.

Mel'cuk, I. 1988. “Semantic description or lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria”. *International Journal of Lexicography* 1 (3), 165–188.

Mel'cuk, I. et al. 1984. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I.* Montreal: University of Montreal Press.

Mel'cuk, I. 1996. "Lexical functions: A tool for the description of lexical relations in the lexicon". En L. Wanner (ed). *Lexical Functions in Lexicology and Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins. 37-102.

Meyer, I. y K. Mackintosh. 1996. "Redefining the terminographer's concept-analysis methods: How can phraseology help?". *Terminology* 3 (1), 1-26.

Miller, G. 1998. “Nouns in WordNet”. En Fellbaum (ed.), *Wordnet: An electronic lexical database*. Cambridge MA: MIT Press, 23-46.

Moreno Ortiz, A. 2000a. “Managing conceptual and terminological information in a user friendly environment”. *Proceedings of OntoLex 2000*. Workshop on Ontologies and Lexical Knowledge Bases. Septiembre 2000, Sofía, Bulgaria.

Moreno Ortiz, A. 2000b. “OntoTerm: un sistema abierto de representación conceptual”. *Actas del XVI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).* Octubre 2000, Vigo, España.

Moreno Ortiz, A. y C. Pérez Hernández. 2000. “Reusing the Mikrokosmos ontology for concept-based multilingual terminology databases”. *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC 2000). Junio 2000, Atenas, Grecia, 1061-1067.

Nelson, K. 1985. *Making Sense: the acquisition of shared meaning*. Orlando: Academic Press.

Pearson, J. 1998. *Terms in Context*. Amsterdam / Philadelphia: John Benjamins.

Pérez Hernández, M.C. 2000. *Explotación de los corpora textuales informatizados para la creación de bases terminológicas basadas en el conocimiento*. Tesis doctoral. Universidad de Málaga.

Peters, C., E. Picchi y L. Biagini. 1996. “Parallel and comparable corpora in language teaching and learning”. En Botley *et al.* (eds.), *Proceedings of Teaching and language Corpora* (UCREL Technical Papers, Volume 9). University of Lancaster, 68-80.

Temmerman, R. 2000. *Towards New Ways of Terminology Desrcription*. Amsterdam / Philadelphia: John Benjamins.

Tercedor Sánchez, M.I. 1999. *La fraseología en el lenguaje biomédico: análisis desde las necesidades del traductor*. Madrid: CSIC / Elies, vol 6. [Disponible en http://elies.rediris.es/elies6/].

Saussure, F. [1916] 1990. *Course de linguistique générale*. París: Payot.

Viegas, E., Mahesh, K., Nirenburg, S. & Beale, S. 1999. “Semantics in Action”. En Saint-Dizier (ed.), *Predicative Forms in Natural Language and Lexical Knowledge Bases*. Dordrecht: Kluwer, 171-204.

Wierzbicka, A. 1992. “In search of tradition: the semantic ideas of Leibniz”. *Lexicographica* 8, 10-25.