ONOMÁZEIN
Revista de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

Diseño e implementación de una base de conocimiento terminológico sobre enfermedades raras

Design and implementation of a terminological knowledge base on rare diseases

Tamara Varela Vila
Universidade de Vigo
España

Elena Sánchez Trigo
Universidade de Vigo
España

ONOMÁZEIN 49 (septiembre de 2020): 01-20
DOI: 10.7764/onomazein.49.01
ISSN: 0718-5758

Tamara Varela Vila: Facultade de Filoloxía e Tradución, Universidade de Vigo, España. | E-mail: tvarela@uvigo.es
Elena Sánchez Trigo: Facultade de Filoloxía e Tradución, Universidade de Vigo, España. | E-mail: etrigo@uvigo.es

Fecha de recepción: abril de 2018
Fecha de aceptación: noviembre de 2018

Resumen

Este trabajo se centra en la elaboración de recursos terminográficos de base conceptual para la gestión de la terminología médica. Se presenta una base de conocimiento terminológico bilingüe (francés y español) sobre un numeroso grupo de enfermedades raras: los errores innatos del metabolismo. El recurso diseñado permite acceder a la terminología utilizada para denominar estas enfermedades, sus síntomas y signos de una forma conceptualmente estructurada. En primer lugar, tras el marco teórico, se describen las fases seguidas para la elaboración de la base de conocimiento: compilación, etiquetado y explotación del corpus de trabajo y construcción de la ontología subyacente. En segundo lugar, se recogen sus datos más significativos y se muestran las funcionalidades de la interfaz de consulta. Su finalidad es describir el subdominio y servir como fuente de información en la traducción especializada.

Palabras clave: base de conocimiento terminológico; ontologías; terminología médica; traducción médica; enfermedades raras.

Abstract

This article focuses on the creation of conceptually-based terminographical resources for managing medical terminology. We present a bilingual (French and Spanish) terminological knowledge base on a large group of rare diseases, namely inborn errors of metabolism. This resource gives its users conceptually structured access to the terminology used when referring to these diseases and their signs and symptoms. After introducing the theoretical framework underpinning its design, we describe the various stages in the creation of the knowledge base itself: compilation, tagging, exploitation of the working corpus and construction of the underlying ontology. We then give an overview of the salient data and provide a guide to the functionalities of the query interface. The purpose of the resource is to describe the sub-domain in question and serve as a source of information for specialist translators.

Keywords: terminological knowledge base; ontologies; medical terminology; medical translation; rare diseases.

1. Introducción[1]

En la actualidad, los proyectos sobre gestión terminológica buscan la creación de recursos que describan satisfactoriamente el ámbito de especialidad en el que se centran y que permitan acceder a la información de una manera dinámica y multidimensional. Esta perspectiva ha conducido al uso de ontologías como elemento subyacente a las bases de conocimiento terminológico. El interés que han suscitado en este campo se basa en su utilidad para formalizar dominios del conocimiento y establecer relaciones entre conceptos y unidades terminológicas (Durán Muñoz y Bautista Zambrana, 2013).

El recurso que presentamos en este trabajo, ONTERMET, se inscribe en esta línea. Se trata de una base de conocimiento terminológico bilingüe, en francés y español, sobre el campo médico de los errores innatos del metabolismo (EIM), un amplio y complejo grupo de enfermedades raras (ER). Tras el marco teórico en el que se inscribe la investigación, se describe la metodología seguida para la creación de la base de conocimiento: compilación de un corpus sobre EIM, proceso de etiquetado y explotación, y construcción de la ontología que formaliza el conocimiento extraído del corpus. A continuación, se recogen las principales características de ONTERMET y se muestran las funcionalidades de la interfaz creada para su consulta.

La investigación se centra en las ER por tratarse de un subdominio de especial interés tanto desde el punto de vista médico como social. Se han descrito más de 6000 enfermedades que afectan, en su conjunto, al 8-10 % de la población de la UE, es decir, unos 30 millones de personas (EURORDIS, 2015). Este número tan importante de afectados hace que constituyan una prioridad en la salud pública europea. Nuestra actividad traductora nos ha permitido constatar la necesidad de desarrollar más recursos terminográficos que describan este subdominio médico de una forma adecuada para resolver las posibles dudas tanto terminológicas como conceptuales que surgen en la realización de una traducción.

La amplitud del campo de las ER, al que nos hemos referido, ha obligado a centrar el trabajo en un grupo específico. Se han seleccionado los EIM porque tanto su elevado número (más de 500) como su heterogeneidad clínica los convierten en un ejemplo representativo del conjunto de ER. Se trata de enfermedades producidas por la mutación o alteración en la secuencia del ADN de un solo gen (monogénicas) y, en general, graves, que causan una afectación multiorgánica.

1 Este trabajo ha sido realizado en el marco de los proyectos de investigación FFI2014-51978-C2-1-R y TIN2017-85160-C2-2-R del Ministerio español de Economía y Competitividad.

2. Marco teórico

Conocer el funcionamiento de la terminología en un ámbito de especialidad constituye un elemento esencial para comprender los mecanismos utilizados por las lenguas para transmitir la información. Los enfoques terminológicos actuales (ej. Gaudin, 2003; Temmerman, 2000; Cabré, 1999) se caracterizan por adoptar una perspectiva integradora, que describe la comunicación especializada atendiendo tanto a los aspectos lingüísticos como cognitivos y comunicativos del concepto. Todos ellos otorgan más importancia al análisis discursivo y, por lo tanto, al uso real de los términos.

Este análisis discursivo se efectúa principalmente empleando la lingüística de corpus (EAGLES, 1996; McEnery, 2003; Sinclair, 2004; Tognini-Bonelli, 2010) por posibilitar el estudio de la lengua a partir de datos reales. De este modo, se ha desarrollado una terminología basada en el análisis de textos de especialidad que permite no solo disponer de material adecuadamente seleccionado para la descripción de los diferentes fenómenos, sino también realizar generalizaciones que el recurso a la intuición de forma única no permitía. Asimismo, la aplicación de nuevas modalidades de representación del conocimiento y la disponibilidad de herramientas para el desarrollo de organizaciones conceptuales más complejas han originado un cambio de rumbo en la práctica terminográfica.

Concretamente, la combinación de las técnicas propuestas por la lingüística de corpus y la terminología descriptiva basada en el conocimiento ha dado lugar a un nuevo enfoque: la 'terminología basada en la representación del conocimiento' (Meyer y Mackintosh, 1996) u 'ontoterminografía' (Durán Muñoz, 2012), entre otras denominaciones[2]. En esta orientación, los corpus lingüísticos sirven como recurso básico para la extracción de información de los textos de especialidad, con el fin de modelar la estructura conceptual de los dominios en cuestión, la cual se ve reflejada en los textos. Esta estructura se representa por medio de distintos formalismos propuestos por la inteligencia artificial, especialmente mediante ontologías (Gruber, 1993; Swartout y otros, 1997; Studer y otros, 1998).

El término 'ontología', proveniente del campo de la filosofía, comienza a utilizarse a partir de los años 90 en el ámbito de las ciencias de la computación y la inteligencia artificial para hacer referencia a recursos que permiten representar el conocimiento compartido sobre un dominio. En concreto, se trata de la herramienta fundamental para organizar el conocimiento o parcelas de este, con la finalidad de facilitar el intercambio de información entre el ser humano y los ordenadores. Desde esta perspectiva, las ontologías son el resultado del análisis y el modelaje ontológico, más que una disciplina (Montiel-Ponsoda, 2009: 153). Su protagonis-

2 Otros autores han denominado a esta nueva línea de investigación 'termontografía' (Temmerman y Kerremans, 2003) o bien 'ontoterminología' (Roche, 2012).

mo actual se debe a la posibilidad de utilizarlas en aplicaciones relacionadas con la gestión del conocimiento, el procesamiento del lenguaje natural, la recuperación de la información, etc. En el ámbito de la terminología, las ontologías constituyen una manera de representar formalmente los conceptos pertenecientes a un ámbito de especialidad y sus relaciones de forma independiente a la lengua, de modo que a esta representación se pueda vincular la terminología en una o más lenguas naturales.

La terminografía basada en la representación del conocimiento ha supuesto una importante mejora en el modo de almacenar, organizar y recuperar la información de los campos especializados, lo que es imprescindible para el correcto uso de su terminología. Dentro de esta línea han surgido diferentes proyectos en los que se recurre a las ontologías para la representación conceptual. En concreto, en el dominio médico cabe mencionar el proyecto GALEN para la clasificación de procedimientos quirúrgicos (Rector y Rogers, 2006); SNOMED CT (International Health Terminology Standards Development Organisation, 2017), terminología médica integral para codificar, recuperar, comunicar y analizar datos médicos; o Disease Ontology (DO) (Bello y otros, 2018). Específicamente en el ámbito español, se pueden citar GENOMA-KB (Cabré y otros, 2004), una base de conocimiento del genoma humano, u OncoTerm, un sistema bilingüe de información y recursos oncológicos (López Rodríguez y otros, 2006).

Todos estos proyectos, cuyas características se contrastarán con la base de conocimiento terminológico ONTERMET en el apartado de conclusiones, tienen en común la búsqueda de sistemas eficaces para representar los ámbitos de especialidad de la medicina en los que se centran y para vincular en estas representaciones cada concepto con los términos utilizados para denominarlos en una o varias lenguas.

Nuestro trabajo se inscribe en esta línea. Utilizamos la lingüística de corpus como metodología para estudiar la dinámica de los términos y extraer información terminológica y conceptual de un corpus de especialidad sobre el dominio de los EIM. La información se formaliza por medio de una ontología que, además de plasmar la situación que ocupa cada uno de los conceptos dentro del sistema conceptual y sus relaciones con otros conceptos, recoge la terminología asociada a cada uno de ellos.

3. Diseño de la base de conocimiento terminológico

El proceso de creación de la base de conocimiento terminológico ONTERMET se realizó en dos fases principales. En primer lugar, se llevó a cabo la compilación del corpus sobre EIM, así como su etiquetado y explotación (extracción de información terminológica y conceptual). En segundo lugar, se construyó la ontología, que se utilizó como «skeletal foundation» (Swartout y otros, 1997: 138) de la base de conocimiento. En los apartados que siguen se presentan los aspectos más destacados en relación con estas diferentes fases.

3.1. Fase 1: compilación, etiquetado y explotación del corpus de trabajo

Los datos que se recogen en ONTERMET fueron obtenidos utilizando como metodología principal la lingüística de corpus. Para ello se creó EMCOR (corpus de enfermedades metabólicas), de donde se extrajo la información conceptual y terminológica del ámbito objeto de estudio[3].

De acuerdo con la clasificación realizada por Laviosa (2002: 33-42), EMCOR es un corpus de textos completos y escritos, sincrónico (publicados entre 2001 y 2013), terminológico o especializado, bilingüe (francés y español) y comparable. Está organizado en dos subcorpus establecidos en función de las lenguas que lo integran y con muestras textuales seleccionadas de acuerdo con los mismos criterios, lo que garantiza su comparabilidad.

En relación con su tamaño, se trata de un corpus mediano (Sardinha, 2002: 119), ya que está formado por un total de 771 223 palabras y 41 043 formas que proceden de 208 muestras textuales incluidas en francés y 203 en español. En la tabla 1 podemos observar los datos estadísticos más significativos de cada uno de estos subcorpus:

TABLA 1
Corpus EMCOR. Datos estadísticos

	SUBCORPUS EN ESPAÑOL	SUBCORPUS EN FRANCÉS
Tamaño (*bytes*)	2 611 049	2 710 154
N.º de palabras	387 706	383 517
N.º de formas	20 048	20 995
Ratio formas/palabras	5.39	5.72
Ratio formas/palabras estandarizada	40.34	41.57
Tamaño medio palabras	5.40	5.62
N.º de oraciones	15 391	16 620
Tamaño medio oraciones	24.16	22.09

3 El corpus EMCOR toma como punto de partida uno más reducido creado para un trabajo de investigación previo con la finalidad de elaborar un glosario sobre EIM (Varela Vila y Sánchez Trigo, 2012). Este corpus inicial se amplió y se modificó a fin de adaptarlo a los nuevos objetivos investigadores. El número de palabras se incrementó en un 68 % y el número de formas en un 34 %, por lo que refleja con mayor precisión el ámbito de los EIM.

Dado que se trata de un corpus especializado y estos, de acuerdo con Bowker y Pearson (2002: 48), tienden a ser más pequeños que los generales, su tamaño es suficiente para representar el subdominio objeto de estudio.

Los géneros textuales representados en EMCOR se seleccionaron siguiendo dos criterios. Por un lado, las propuestas sobre géneros médicos realizadas por Montalt y González Davies (2007), Posteguillo Gómez y Piqué-Angordans (2007) y García Izquierdo (2009). Por otro, realizando un análisis de los géneros textuales más representativos dentro del ámbito de los EIM. De este modo, se optó por incluir en el corpus: artículo original, artículo de revisión, caso clínico, resumen de artículo científico e información para pacientes. Los cuatro primeros géneros, mayoritarios en el corpus, son especializados (90,2 %) y, por su parte, información para pacientes (9,8 %) puede clasificarse como semiespecializado o divulgativo, dependiendo de su procedencia. Se ha seguido en este punto a Vargas Sierra (2006: 6), que considera que es interesante incluir esta gradación para que estén representados los diferentes tipos de datos lingüísticos. La tabla 2 muestra el número de palabras por género textual que presenta EMCOR:

TABLA 2
Corpus EMCOR. Palabras por género textual

GÉNEROS TEXTUALES	N.º PALABRAS ESPAÑOL	N.º PALABRAS FRANCÉS
Artículo de revisión	184 201	184 274
Artículo original	48 370	52 252
Caso clínico	113 494	103 069
Información para pacientes	37 342	38 000
Resumen de artículo científico	4299	5922
TOTALES	**387 706**	**383 517**

Una vez compilado el corpus, se etiquetó morfosintácticamente utilizando el programa TreeTagger. Se trata de uno de los etiquetadores más empleados, que permite etiquetar textos, entre otras lenguas, en francés y en español.

La extracción de información terminológica y conceptual de EMCOR se realizó de modo semiautomático, por medio del conjunto de herramientas informáticas WordSmith Tools. El análisis se dividió en cuatro etapas principales, que nos permitieron obtener una visión global de la terminología perteneciente al campo de los EIM y entender las relaciones que se establecen en él. Se trata de las siguientes etapas: a) análisis estadístico, b) análisis de las palabras más frecuentes, c) análisis de los patrones sintácticos y d) análisis de los marcadores lingüísticos de relación conceptual. Posteriormente, se efectuaron asimismo múltiples con-

sultas puntuales para obtener información concreta que permitiese extraer toda la terminología necesaria y determinar las relaciones entre los distintos conceptos.

En el campo de la medicina existe una gran diversidad de este tipo de relaciones, lo que ha obligado a realizar una selección utilizando para ello el criterio de mayor representatividad en la descripción del subdominio. El análisis de EMCOR se centró, por lo tanto, en identificar los términos utilizados para denominar las enfermedades que forman parte del grupo de los EIM y sus síntomas y signos, así como en determinar las relaciones que se establecen entre ellos.

Si bien del corpus se extrajo la información principal para crear ONTERMET, para validarla y complementarla se recurrió, asimismo, a expertos del Instituto de Investigación en Enfermedades Raras (IIER) del Instituto de Salud Carlos III (ISCIII) de Madrid.

3.2. Fase 2: creación de la ontología

La base de conocimiento terminológico está conformada por una ontología subyacente, cuya finalidad es estructurar el conocimiento que en ella se recoge. Se trata, en concreto, de una ontología de dominio, ya que formaliza un ámbito específico (Mizoguchi y otros, 1995; Heijst y otros, 1997), en este caso, el de los EIM. Para su creación se utilizaron los siguientes recursos: el editor de ontologías Protégé[4]; el modelo de datos SKOS (Simple Knowledge Organization System)[5]; y la ontología de relaciones OBO[6].

La ontología creada almacena de manera jerárquica los grupos de conceptos seleccionados: las enfermedades (EIM) y sus síntomas o signos. Para ello, se tomaron como base organizaciones conceptuales aceptadas dentro del ámbito de especialidad: Orphanet (portal de referencia sobre enfermedades raras y medicamentos huérfanos), para clasificar los EIM, y la CIE-10 (Clasificación Internacional de Enfermedades), para clasificar los signos y síntomas.

Del mismo modo, en la ontología se formalizan las relaciones existentes entre los conceptos. Esto es, qué síntomas o signos afectan a cada enfermedad y, a la inversa, qué enfermedades pueden presentar determinado síntoma o signo.

En la figura 1 se muestra, como ejemplo, la formalización de los EIM. El punto de partida es el concepto 'enfermedad', con un único subtipo (el concepto 'error innato del metabolismo').

4 Protégé (http://protege.stanford.edu) es el editor más utilizado en estos momentos. Constituye el estándar *de facto* dentro del conjunto de los editores en OWL (el lenguaje estándar de representación del conocimiento y utilizado para crear la ontología de este trabajo).

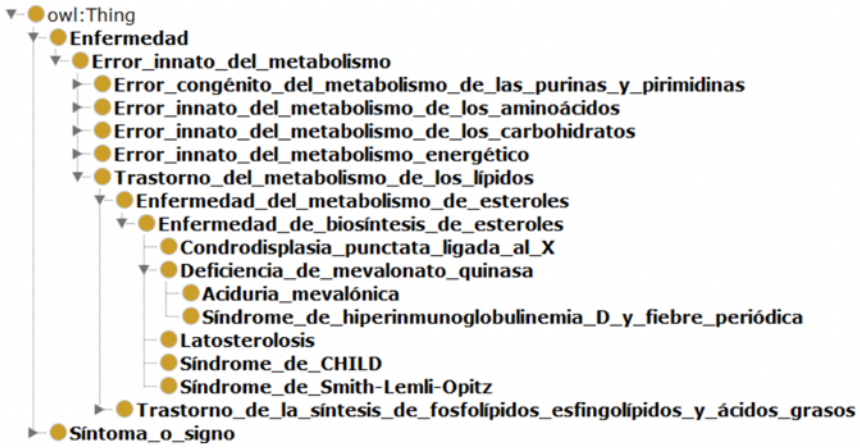
5 SKOS (https://www.w3.org/2004/02/skos) es un modelo de datos para compartir y vincular vocabularios controlados a través de la web.

6 OBO (http://obofoundry.org/ontology/ro.html) es una ontología que formaliza un conjunto de relaciones rigurosamente definidas que se utilizan habitualmente en las ontologías biomédicas.

Este, a su vez, está subdividido en los cinco grandes grupos de EIM existentes, que se van concretando en los niveles inferiores:

FIGURA 1

Formalización de los EIM en la ontología



La organización creada para representar el concepto 'síntoma o signo' consta, por su parte, de 19 grandes grupos que se van subdividiendo hasta alcanzar el concepto más concreto.

Una vez creadas estas organizaciones jerárquicas, se atribuyó a cada concepto los términos empleados para denominarlo. Se utilizaron las etiquetas «skos:prefLabel» (etiqueta preferida) y «skos:altLabel» (etiqueta alternativa), que pertenecen al modelo de datos SKOS ya mencionado. Cada concepto tiene un término preferido por lengua, pero se recogen también las variantes denominativas, como se puede observar en la siguiente figura:

FIGURA 2

Etiquetas preferidas y alternativas vinculadas a un solo concepto

El hecho de que los conceptos se separen claramente de sus denominaciones hace que la ontología sea independiente de la lengua de trabajo y permite, asignando a cada concepto etiquetas denominativas en otros idiomas, una futura ampliación de la base de conocimiento sin tener que llevar a cabo todos los pasos realizados inicialmente para crearla.

La vinculación realizada de cada uno de los conceptos con su código CIE-10 (mediante la etiqueta «oboInOwl:hasDbXref» de la ontología de relaciones OBO) permite que ONTERMET contenga vínculos con la Clasificación Internacional de Enfermedades a partir de cada uno de los conceptos consultados. La etiqueta utilizada en este caso también podrá utilizarse en un futuro para ampliar el trabajo y crear referencias cruzadas con otras bases de datos, como Orphanet, a la que ya nos hemos referido, u OMIM (Online Mendelian Inheritance in Man), base de datos de enfermedades genéticas del ser humano, de interés por tener los EIM este origen.

En aquellos conceptos en los que se consideró necesario hacer alguna recomendación sobre la terminología usada para denominarlos, se introdujeron notas (etiqueta «rdfs:comment»). Las recomendaciones se basan en la información aportada por obras de referencia en el campo de la medicina, así como por los especialistas que colaboraron en la revisión de esta base de conocimiento.

Una vez introducidos en la ontología todos los conceptos y asignadas las etiquetas correspondientes a las que nos hemos referido, se procedió a crear los vínculos entre los EIM y sus síntomas y signos. Para llevar esto a cabo, fue necesario crear dos propiedades de objetos inversas, que permitiesen enlazar unos conceptos con otros: «hasSymptom» (tiene síntoma) y «isSymptomOf» (es síntoma de). De este modo, partiendo de la información obtenida en el corpus EMCOR, se vincularon los distintos EIM con sus posibles síntomas y signos.

La ontología descrita formaliza la información extraída del corpus EMCOR y constituye el núcleo que estructura la base de conocimiento sobre EIM.

4. Base de conocimiento terminológico ONTERMET

En ONTERMET están recogidos un total de 2124 conceptos que pertenecen a uno de los siguientes grupos: aquellos que representan enfermedades, es decir, el conjunto de enfermedades que integran los EIM (185 conceptos); aquellos que representan síntomas o signos (1944 conceptos); y aquellos que pueden constituir tanto enfermedades como síntomas o signos (5 conceptos).

Si bien algunos de estos conceptos —procedentes, como ya indicamos, de organizaciones conceptuales aceptadas dentro del ámbito de especialidad— no figuraban en el corpus EMCOR, se han mantenido porque sirven para aportar un mayor grado de especificidad a la

organización conceptual del subdominio[7]. De acuerdo con esto, de los 185 conceptos referidos a los EIM, 167 están recogidos en EMCOR. Por lo que respecta a los 1944 conceptos que representan síntomas o signos, 1026 aparecen recogidos en EMCOR.

En relación con la información terminológica, la base de conocimiento presenta un total de 1188 términos preferidos en español y en francés (lo que coincide, como es lógico, con el número de conceptos que figuran en EMCOR, esto es, un término preferido por concepto y lengua). Asimismo, recoge 1357 denominaciones alternativas en español y 901 en francés. Esta diferencia evidencia que el grado de variación terminológica en español es superior al del francés en el campo de los EIM.

Si bien ONTERMET constituye una primera aproximación por nuestra parte a la representación del amplio campo de las ER, los datos presentados muestran la representatividad de la información que se recoge en la base de conocimiento.

4.1. Interfaz de consulta

ONTERMET puede consultarse en línea por medio de una interfaz en la dirección www.ontermet.com. Al acceder, el usuario obtiene una primera visualización en francés o en español, según la configuración de su navegador. Es posible cambiar el idioma de visualización (extremo superior derecho de la pantalla), lo que es esencial para los traductores, pues facilita la obtención de los términos equivalentes en una y otra lengua de trabajo.

El acceso a los datos puede realizarse de dos formas distintas: a través de la organización conceptual que aparece en la interfaz a la izquierda de la pantalla (enfoque onomasiológico) o utilizando la búsqueda por palabra clave en la caja situada en la parte derecha de la barra superior (enfoque semasiológico).

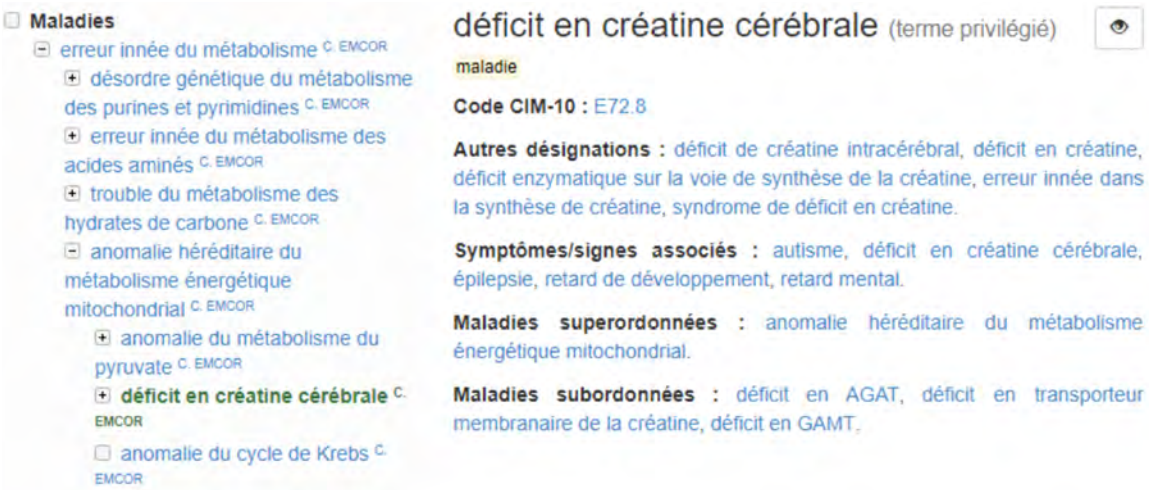
En la figura 3 se presenta cómo se visualiza la información sobre un concepto en ONTERMET, en este caso ‘déficit en créatine cérébrale’. En la columna izquierda, como acabamos de indicar, se muestra la organización conceptual, mientras que a la derecha se presenta la información que se obtiene sobre el concepto en cuestión, y que explicamos más en detalle en los párrafos que siguen con otros ejemplos.

Los diferentes hipervínculos permiten al usuario navegar por la base de conocimiento para explorar el contenido de ONTERMET de modo intuitivo y no lineal, pasando de unos conceptos a otros relacionados.

7 En la base de conocimiento, los conceptos recogidos en EMCOR figuran con la indicación «C. EMCOR» en superíndice.

FIGURA 3

Información sobre el concepto 'déficit en créatine cérébrale' recogida en ONTERMET



La información específica que se puede consultar relativa a cada uno de los conceptos depende de su naturaleza (enfermedad / síntoma o signo), como se muestra en la tabla 3:

TABLA 3

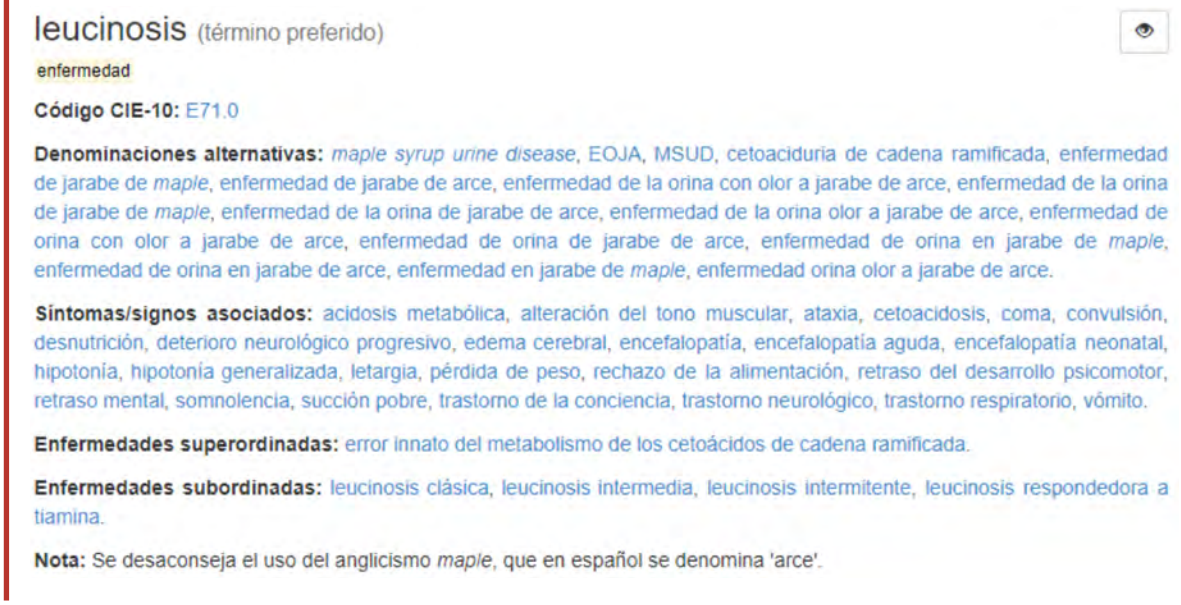
Información recogida en ONTERMET en función del tipo de concepto

ENFERMEDAD	SÍNTOMA O SIGNO
– Término preferido para denominar el concepto	– Término preferido para denominar el concepto
– Tipo de concepto (enfermedad)	– Tipo de concepto (síntoma/signo)
– Código CIE-10	– Código CIE-10
– Denominaciones alternativas del concepto	– Denominaciones alternativas del concepto
– Síntomas/signos asociados	– Enfermedades asociadas
– Enfermedades superordinadas	– Síntomas/signos superordinados
– Enfermedades subordinadas	– Síntomas/signos subordinados
– Nota sobre el uso de alguno de los términos utilizados para denominar el concepto	– Nota sobre el uso de alguno de los términos utilizados para denominar el concepto

A modo de ejemplo, en la figura 4 puede consultarse la información sobre el concepto 'leucinosis' en español (véase la figura en la página siguiente).

Como se puede observar, se indica que 'leucinosis' es la denominación preferida de este concepto en español y que se trata de una enfermedad, no de un síntoma o signo (tipo de concep-

FIGURA 4
Información sobre el concepto 'leucinosis' recogida en ONTERMET



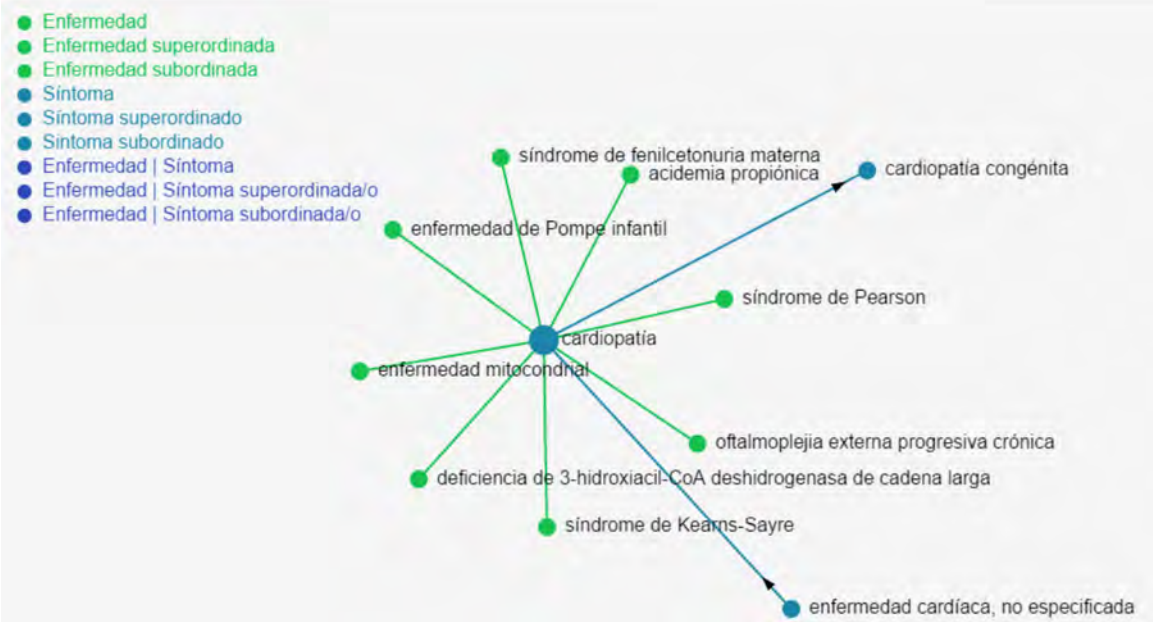
to). A continuación, se incluye el código CIE-10 con su hipervínculo, lo que permite el acceso directo (E71.0, en este caso). Se recogen las denominaciones alternativas de este concepto y los síntomas o signos que puede presentar. Se señala a qué grupo de enfermedades pertenece (enfermedades superordinadas) y qué otras enfermedades agrupa (enfermedades subordinadas). Por último, se incluye una nota que indica que las denominaciones compuestas por el anglicismo *maple* no son recomendables, pues este árbol se denomina en español 'arce'.

Para buscar la información correspondiente a este mismo concepto en francés, solamente habría que seleccionar esta lengua en la parte superior de la pantalla.

Junto a este modo más clásico de visualizar la información, también es posible acceder a la estructura conceptual del ámbito de los EIM en forma de mapa conceptual[8], mediante el icono situado a la derecha del término preferido. El gráfico diseñado permite la consulta de la ontología de un modo más visual y la navegación a través de los diferentes conceptos. Los distintos tipos de conceptos se diferencian mediante un código de colores (enfermedades: verde; síntomas o signos: azul; y enfermedades que pueden ser también síntomas o signos de otros EIM: violeta).

8 Para el diseño de este gráfico interactivo se utilizó D3.js (Bostock y otros, 2011), librería de JavaScript, que permite representar cada uno de los conceptos en formato SVG e incorporar a una web gráficos, datos o infografías interactivas y accesibles.

FIGURA 5
Mapa conceptual de 'cardiopatía'



La figura 5 muestra como ejemplo el concepto 'cardiopatía' (síntoma), que ocupa el lugar central. Los conceptos más cercanos a este son las enfermedades relacionadas (ej. síndrome de Pearson) y los más alejados son los síntomas o signos subordinados y superordinados (ej. cardiopatía congénita). Clicando en cada uno de los nodos de los conceptos es posible navegar por el mapa conceptual y, por lo tanto, consultarlo de un modo más natural e intuitivo.

Las diferentes formas de acceso y visualización de los datos de ONTERMET no son excluyentes, sino que es aconsejable utilizarlas de manera combinada para obtener el máximo rendimiento del recurso.

5. Conclusiones

En este trabajo se han descrito las diferentes etapas seguidas para el diseño e implementación de la base de conocimiento terminológico ONTERMET. Por un lado, se ha explicado el proceso de compilación y explotación del corpus EMCOR. Dado que este constituye la fuente primordial de información conceptual y terminológica, se ha considerado necesario mostrar cómo se ha llevado a cabo. Por otro lado, se han expuesto los pasos seguidos para la creación de la ontología que recoge, de forma conceptualmente estructurada, todo el conocimiento obtenido sobre el campo de los EIM, así como su terminología. Para finalizar, se han presentado los datos más significativos de esta base de conocimiento terminológico, ONTERMET, y se han mostrado las diferentes funcionalidades de la interfaz de consulta.

El proyecto presentado surge de la necesidad de desarrollar recursos terminológicos y conceptuales avanzados que permitan aumentar la calidad de los resultados obtenidos en el proceso de documentación que se lleva a cabo en la traducción especializada. Estos recursos deben estar conformados no solo por información terminológica o lingüística relevante, sino que además deben estar conceptualmente estructurados y adaptarse a las distintas necesidades de consulta de los usuarios.

Como ya se ha indicado en el marco teórico, la línea de investigación en la que se inscribe ONTERMET, las ontologías como sistema de representación del conocimiento médico, ha dado lugar a diferentes proyectos tanto internacionales como en el ámbito español. Podemos citar en primer lugar GALEN (General Architecture for Languages Encyclopædias and Nomenclatures in Medicine), desarrollado para crear métodos y herramientas que permitiesen generar y mantener clasificaciones de procedimientos quirúrgicos[9]. Contiene los conceptos elementales y las relaciones que controlan su combinación para crear otros más complejos. Partiendo de los términos elementales y empleando ciertas reglas gramaticales para combinarlos, construye una expresión en lenguaje natural para cualquier concepto complejo. Un segundo proyecto es SNOMED CT, una terminología clínica integral y multilingüe que puede emplearse para codificar, recuperar, comunicar y analizar datos médicos. Su contenido se representa usando tres tipos de componentes diferentes: conceptos organizados en jerarquías, descripciones que vinculan los diferentes términos en lenguaje natural con los conceptos y relaciones que vinculan cada concepto con otros relacionados. SNOMED CT tiene una versión en inglés y traducciones a distintas lenguas. Como tercer y último ejemplo mencionaremos uno de los recursos de naturaleza ontológica más conocidos en el campo biomédico, Disease Ontology (DO), una ontología estandarizada sobre las enfermedades humanas frecuentes y raras. Su finalidad es proporcionar a la comunidad biomédica descripciones sostenibles, reutilizables y coherentes sobre las enfermedades, sus características fenotípicas y otros conceptos relacionados con estas. Se desarrolló para crear una estructura única que unificase la representación de las enfermedades entre las diversas terminologías y vocabularios por medio de una ontología. Ofrece una definición clara de cada enfermedad dentro de una clasificación basada en su etiología para poder anotar datos biomédicos (Schriml y otros, 2012). La DO integra conceptos y términos extraídos de otras fuentes, como por ejemplo la CIE-9, el National Cancer Institute (NCI) Thesaurus, SNOMED CT, MeSH, OMIM y Orphanet.

La diferencia entre estos proyectos y ONTERMET radica fundamentalmente en la importancia otorgada a la terminología. El principal objetivo de los tres proyectos mencionados es

9 La representación conceptual de estas clasificaciones se lleva a cabo con un lenguaje específico, denominado GRAIL (GALEN Representation and Integration Language), especialmente diseñado para crear las restricciones específicas que se emplean en el ámbito médico.

estructurar, por medio de una ontología, el dominio o subdominio en que se centran. Para ello, se explicitan las cualidades de los diferentes conceptos y se establecen restricciones y reglas de combinación entre ellos. Si bien contienen también terminología asociada a los conceptos, su finalidad no es principalmente terminológica y, por lo tanto, no se le confiere a esta información una importancia central. Esto se manifiesta, por ejemplo, en que la terminología para los conceptos compuestos se genere automáticamente, como sucede en GALEN; en que las distintas terminologías sean traducciones de la versión en inglés, como es el caso de SNOMED CT; o en que se recoja terminología procedente de diferentes recursos previamente existentes, como ocurre en Disease Ontology. Por su parte, ONTERMET, frente a estos, emplea una ontología para estructurar el ámbito de las enfermedades raras con el fin de almacenar de forma organizada la terminología extraída del corpus EMCOR, creado especialmente para servir de base a este recurso. El valor de la ontología creada no solo radica en la organización de este ámbito de conocimiento, sino también en la variedad de términos asociados a cada concepto, que es un reflejo del uso real de la terminología en este campo.

Existen, no obstante, otros proyectos similares a ONTERMET, en los que el uso de ontologías tiene un fin terminológico. Se puede citar, por ejemplo, en el ámbito español, GENOMA-KB, un banco de conocimiento sobre el genoma humano formado por cuatro módulos: corpus textual y base de datos factográfica, terminológica y ontológica. Los dos últimos se construyeron de forma paralela con la herramienta de gestión terminológica denominada OntoTerm, que permite vincular los distintos términos a los conceptos previamente representados en la ontología. Este recurso, no obstante, no está disponible actualmente en línea. Otro proyecto desarrollado en esta misma línea es OncoTerm, una base terminológica que recoge términos en inglés y español extraídos de un corpus especializado y que sirven para denominar conceptos relacionados con el cáncer.

Al igual que ONTERMET, ambos proyectos tienen por objetivo crear sistemas eficaces para representar los ámbitos en los que se centran y enlazar cada concepto con los términos reales usados para denominarlos en una o varias lenguas. A diferencia de estos, sin embargo, ONTERMET se caracteriza por su accesibilidad y facilidad de utilización. Los diferentes modos de consulta disponibles a través de la interfaz diseñada permiten acceder al contenido tanto onomasiológicamente, por medio de la jerarquía de conceptos, como semasiológicamente, empleando la herramienta de búsqueda por término o navegando a través de los distintos hipervínculos que contiene cada entrada. Del mismo modo, se facilita el proceso de traducción entre el francés y el español, pues permite acceder directamente a la terminología equivalente en ambas lenguas.

Junto a las señaladas, otra de las características diferenciadoras de ONTERMET es que se centra en un ámbito de estudio innovador y de actualidad, como es el de las ER, con el objetivo de crear recursos que las describan de forma adecuada. Por otra parte, es también de especial interés el par de lenguas seleccionado en un campo como el de la medicina, en

que el uso del inglés como *lingua franca* es predominante. En este sentido, consideramos importante poder contar con recursos que describan adecuadamente la terminología médica en francés y español.

Para finalizar, cabe señalar que la base de conocimiento terminológico creada constituye una primera aproximación por nuestra parte al desarrollo de herramientas multilingüísticas, multidimensionales y dinámicas que faciliten el trabajo de los traductores especializados y reduzcan la necesidad de consulta de diferentes fuentes especializadas. Se trata, asimismo, de un primer modelo fácilmente ampliable. En efecto, una de las posibles líneas de trabajo en el futuro es su extensión a todo el campo de las ER, así como la ampliación del catálogo de relaciones conceptuales representadas, ya que su diversificación permitirá una mejor descripción del ámbito.

6. Bibliografía citada

Bello, Susan, y otros, 2018: "Disease Ontology: improving and unifying disease annotations across species", *Disease Models & Mechanisms* 11 [https://goo.gl/ydtqvf, fecha de consulta: 22 de noviembre de 2018].

Bostock, Michael, Vadim Ogievetsky y Jeffrey Heer, 2011: "D^3: Data-Driven Documents", *ieee Transactions on Visualization and Computer Graphics* 17, 12 [http://goo.gl/QnJPMj, fecha de consulta: 22 de noviembre de 2018].

Bowker, Lynne, y Jennifer Pearson, 2002: *Working with specialized language: a practical guide to using corpora*, Londres: Routledge.

Cabré, María Teresa, y otros, 2004: "The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities" en *LREC 2004. Fourth International Conference on Language Resources and Evaluation*, Lisboa: European Languages Resources Association, 87-90 [https://goo.gl/wJ2EYF, fecha de consulta: 22 de noviembre de 2018].

Cabré, María Teresa, 1999: *La Terminología: Representación y Comunicación. Elementos para una teoría de base comunicativa y otros artículos*, Barcelona: IULA, Universitat Pompeu Fabra.

Durán Muñoz, Isabel, y María Rosario Bautista Zambrana, 2013: "Applying Ontologies to Terminology: Advantages and Disadvantages", *Hermes - Journal of Language and Communication in Business* 51, 65-77 [https://goo.gl/DGXcxR, fecha de consulta: 22 de noviembre de 2018].

Durán Muñoz, Isabel, 2012: *La ontoterminografía aplicada a la traducción*, Frankfurt: Peter Lang.

EAGLES, 1996: *Preliminary Recommendations on Corpus Typology*, EAG-TCWG-CTYP/P [http://goo.gl/ZR17xc, fecha de consulta: 22 de noviembre de 2018].

EURORDIS, 2015: "Sobre las Enfermedades Raras" [https://www.eurordis.org/es/enfermedades-raras, fecha de consulta: 22 de noviembre de 2018].

García Izquierdo, Isabel, 2009: *Divulgación médica y traducción. El género Información para pacientes*, Berna: Peter Lang.

Gaudin, François, 2003: *Socioterminologie. Une approche sociolinguistique de la terminologie*, Bruselas: De Boeck/Duculot.

Gruber, Tom, 1993: "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition* 5, 2, 199-220.

Heijst, Gertjan van, August Schreiber y Bob Wielinga, 1997: "Using explicit ontologies in KBS development", *International Journal of Human-Computer Studies* 46, 2-3, 183-292 [http://goo.gl/Q3nUYE, fecha de consulta: 22 de noviembre de 2018].

International Health Terminology Standards Development Organisation, 2017: *SNOMED CT Starter Guide* [https://goo.gl/LGcuZ6, fecha de consulta: 22 de noviembre de 2018].

Laviosa, Sara, 2002: *Corpus-based translations studies: theory findings applications*, Ámsterdam: Rodopi.

López Rodríguez, Clara Inés, Pamela Faber y María Isabel Tercedor Sánchez, 2006: "Terminología basada en el conocimiento para la traducción y la divulgación médicas: el caso de Oncoterm", *Panace@* 7, 24, 228-240 [http://goo.gl/619Gj7, fecha de consulta: 22 de noviembre de 2018].

McEnery, Tony, 2003: "Corpus linguistics" en Ruslan Mitkov (ed.): *Oxford handbook of computational linguistics*, Oxford: Oxford University Press, 448-463.

Meyer, Ingrid, y Kristen Mackintosh, 1996: "The Corpus from a Terminographer's Viewpoint", *International Journal of Corpus Linguistics* 1, 2, 257-268.

Mizoguchi, Riichiro, Johan Vanwelkenhuysen y Mitsuru Ikeda, 1995: "Task Ontology for reuse of problem solving knowledge" en Nicolaas Mars (ed.): *Towards very large knowledge bases: knowledge building and knowledge sharing*, Ámsterdam: IOS Press, 46-57.

Montalt Resurrecció, Vicent, y María González Davies, 2007: *Medical translation step by step: learning by drafting*, Manchester: St. Jerome Publishing.

Montiel-Ponsoda, Elena, 2009: "Ontology Localization: a Key Issue in the Semantic Web of the Future" en Gred Wotjak, Vessela Ivanova y Encarnación Tabares Plasencia (eds.): *Translatione via*

facienda. Festschrift für Christiane Nord zum 65. Geburtstag. Homenaje a Christiane Nord en su 65 cumpleaños, Frankfurt: Peter Lang, 153-167.

Posteguillo Gómez, Santiago, y Jordi Piqué-Angordans, 2007: "El lenguaje de las ciencias médicas: comunicación escrita" en Enrique Alcaraz Varó, José Mateo Martínez y Francisco Yus Ramos (eds.): *Las lenguas profesionales y académicas*, Barcelona: Ariel, 167-190.

Rector, Alan, y Jeremy Rogers, 2006: "Ontological and practical issues in using a description logic to represent medical concept systems: Experience from GALEN" en Pedro Barahona y otros (eds.): *Reasoning Web. Second International Summer School 2006, Lisbon, Portugal, September 4-8, 2006, Tutorial Lectures*, Berlín/Heidelberg: Springer, 197-231.

Roche, Christophe, 2012: "Ontoterminology: How to unify terminology and ontology into a single paradigm" en *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, Estambul: European Language Resources Association, 2626-2630 [http://goo.gl/bYaGHx, fecha de consulta: 22 de noviembre de 2018].

Sardinha, Tony, 2002: "Tamanho de corpus", *The ESPecialist* 23, 2, 103-122 [https://goo.gl/iqTCtG, fecha de consulta: 22 de noviembre de 2018].

Schriml, Lynn Marie, y otros, 2012: "Disease Ontology: a backbone for disease semantic integration", *Nucleic Acids Research* 40, D940-D946 [http://goo.gl/TMYR02, fecha de consulta: 22 de noviembre de 2018].

Sinclair, John, 2004: "Corpus and Text: Basic Principles" en Martin Wynne (ed.): *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books [http://goo.gl/uixXjh, fecha de consulta: 22 de noviembre de 2018].

Studer, Rudi, Richard Benjamins y Dieter Fensel, 1998: "Knowledge Engineering: Principles and Methods", *Data & Knowledge Engineering* 25, 1-2, 161-197.

Swartout, Bill, y otros, 1997: "Toward Distributed Use of Large-Scale Ontologies" en Adam Farquhar y Michael Gruninger (eds.): *Papers from the 1997 AAAI Spring Symposium*, California: The AAAI Press, 138-148.

Temmerman, Rita, y Koen Kerremans, 2003: "Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description" en Eva Hajičová, Anna Kotěšovcová y Jiří Mírovský (eds.): *Proceedings of CIL17*, Praga: Matfyzpress, MFF UK.

Temmerman, Rita, 2000: *Towards new ways of terminology description: the sociocognitive approach*, Ámsterdam/Filadelfia: John Benjamins.

Tognini-Bonelli, Elena, 2010: "Theoretical overview of the evolution of corpus linguistics" en Anne O'Keeffe y Michael McCarthy (eds.): *The Routledge Handbook of Corpus Linguistics*, Londres: Routledge, 14-27.

Varela Vila, Tamara, y Elena Sánchez Trigo, 2012: "EMCOR: a corpus for terminological medical purposes", *JoSTrans: Journal of Specialised Translation (Special Issue on Terminology, Phraseology and Translation)* 18, 139-159 [http://goo.gl/5ElwVP, fecha de consulta: 22 de noviembre de 2018].

Vargas Sierra, Chelo, 2006: "Diseño de un corpus especializado con fines terminográficos: el Corpus de la Piedra Natural", *Debate Terminológico* 2, 7, París: RITERM (Red Iberoamericana de Terminología) [https://goo.gl/ZeL1Qq, fecha de consulta: 22 de noviembre de 2018].