LSP Professional Communication Knowledge Management Cognition
LSP Journal, Vol.4, No.1 (2013) / http://lsp.cbs.dk

Traducción de textos biomédicos: creación de recursos a partir de un corpus sobre enfermedades neuromusculares pediátricas (francés-español)

[a]Elena Sánchez Trigo, [b]María Magdalena Vila Barbosa

[a]Departamento de Traducción y Lingüística
Universidade de Vigo
Vigo, España
etrigo@uvigo.es

[b]Departamento de Traducción y Lingüística
Universidade de Vigo
Vigo, España
mariamagdavilabarbosa@uvigo.es

Keywords: *translation resources, specialized corpora, rare diseases, medical texts, neuromuscular diseases, terminological glossary*

Abstract

This paper deals with the creation of terminology resources to assist in the translation of biomedical texts. First of all, we describe the criteria used to design and compile the ENEUPECOR corpus, a French-Spanish bilingual specialised corpus constituted by scientific papers on neuromuscular diseases in paediatrics. Afterwards, we describe the methodology and steps involved in corpus exploitation in order to create a French-Spanish bilingual glossary including the main concepts of the sub-domain selected.

This study is part of our current research on text translation within the field of rare diseases, domain to which belong neuromuscular diseases in paediatrics. Both the thematic sub-domain selected as well as the choice of languages constitute a novel line of research. Furthermore, the fact that rare diseases are now beginning to be seen as a priority in European Public Health Policies confers more relevance to our work from a social point of view. Thus, academic-scientific interests and social interests go hand-in-hand.

1 Introducción[1]

La pericia o conocimiento experto (Sirén & Hakkrarainen, 2002; Shreve, 2002) de un traductor de textos del ámbito de la biomedicina debe incluir, entre otros aspectos, un buen conocimiento de la terminología propia del subdominio en el que se inscriben los textos que

[1] Este artículo forma parte de los trabajos realizados con la financiación de la *Consellería de Cultura, Educación e Ordenación Universitaria* (Xunta de Galicia- España) a través de las ayudas para la *Consolidación y Estructuración de Unidades de Investigación Competitivas del Sistema Universitario de Galicia* (refs. CN2012/317 CN2012/319 y CN2012/259) y por la *Consellería de Economía e Industria* (Xunta de Galicia-España) a través del Plan I2C (2011-2015).

traduce. La gran diversidad de especialidades que componen este ámbito, la rápida evolución de las mismas y la constante creación de nuevas microáreas hacen que la disponibilidad de recursos terminográficos constituya una herramienta fundamental para realizar su trabajo.

La investigación sobre corpus aplicada a la traducción, denominada en inglés *Corpus Translations Studies* (CTS), se remonta a principios de los años noventa del siglo XX. En 1993 Mona Baker sostenía que las técnicas estadísticas y la metodología puestas en funcionamiento por Sinclair (1991) podían contribuir al paso de estudios prescriptivos a descriptivos. Apuntaba, asimismo, que el cambio de orientación desde una perspectiva conceptual a un enfoque situacional y del significado al uso ponía las bases del desarrollo de una nueva manera de trabajar que tomaba el texto como centro de estudio. Más recientemente, Bernardini, Stewart y Zanettin (2003) denominan *Applied Corpus-Based Translation Studies* a la rama resultante de la convergencia de CTS y *Applied Translation Studies*.

La aplicación del trabajo con córpora a la traducción especializada permite obtener información fundamental y de diverso tipo sobre los textos:

> A corpus can be a useful resource for learning about the linguistic features of an LSP, such as knowledge about terms, collocations, grammar and style. It can also provide conceptual information, such as knowledge about the characteristics of the concepts behind the terms and about the relationships concepts have with one another (Bowker & Pearson, 2002: 39).

En la misma medida en que existe un cierto convencionalismo en la lengua común, que caracteriza la forma peculiar de expresión de una comunidad lingüística, en los lenguajes de especialidad se reconocen diversos convencionalismos que caracterizan la forma de decir las cosas y de organizar la información en una determinada área del conocimiento (Tagnin, 2005). El acceso a esta información permitirá al traductor producir textos de calidad que no resulten extraños a los lectores profesionales y que encajen perfectamente con los textos originales producidos en esa área.

El estudio de extensas colecciones de textos, gracias al importante volumen de textos accesibles en la red, se ha revelado como un procedimiento útil para la creación de recursos terminográficos multilingües. La extracción de conocimiento mediante corpus multilingües permite estudiar las lenguas de especialidad en el uso que de ellas hacen sus productores naturales o quienes las emplean para comunicarse (Pérez, 2002).

De acuerdo con estos planteamientos, el objetivo de este artículo es presentar un ejemplo de las posibilidades que la explotación de un corpus de especialidad ofrece a la práctica terminográfica y a los estudios de traducción. En primer lugar abordamos los pasos seguidos para la creación del corpus ENEUPECOR. Un corpus bilingüe (francés y español) especializado en un subdominio de la biomedicina (enfermedades neuromusculares (ENM) pediátricas) y en un género textual específico (artículo científico). A continuación describimos la metodología y herramientas utilizadas para llevar a cabo su explotación con la finalidad de elaborar un glosario en francés y español sobre el ámbito indicado. Se explora, por ejemplo, el interés de algunas herramientas específicas como el extractor de contextos definitorios ECODE. El artículo se cierra con las principales conclusiones obtenidas tras el estudio realizado.

En estos momentos las enfermedades raras (ER), ámbito en el que se integran las ENM pediátricas que conforman ENEUPECOR, empiezan a constituir una prioridad en las políticas de salud pública europeas. Así, el 2013 ha sido declarado en el año español de las ER. La escasez de recursos de utilidad para traductores centrados en estas enfermedades nos ha llevado a desarrollar una línea de investigación para contribuir a su creación (Miquel & Sánchez, 2010; Varela & Sánchez, 2012).

Tanto la actualidad social del subdominio temático, que no ha sido abordado, como la metodología y las lenguas en la que nos centramos confieren elementos novedosos al trabajo que presentamos.

2 El corpus ENEUPECOR: subdominio temático, criterios de diseño y datos estadísticos

Como hemos indicado, las ENM pediátricas se integran en el ámbito de las denominadas ER, también conocidas como 'minoritarias', 'huérfanas' o 'poco frecuentes'. Esta denominación engloba a un amplio número de formas de expresión, aproximadamente unas 200, consideradas de baja prevalencia, pero que incluyen a un importante número global de personas afectadas (de 24 a 36 millones de personas en la UE).

La selección temática de nuestro corpus y de esta línea de investigación se justifica inicialmente por nuestra experiencia en la traducción de textos de este ámbito. De este modo hemos podido conocer la existencia de una demanda de traducciones que se deriva de la actual búsqueda de visibilización de las ER. Esta actividad nos ha permitido, asimismo, profundizar en las características de este conjunto de afecciones y en la producción textual sobre las mismas.

El hecho de centrarnos en esta ocasión en el segmento específico de las ENM pediátricas se justifica porque la mayoría de las ENM se manifiestan en el período neonatal y en los primeros años de vida. Un diagnóstico temprano de las ENM va a ser, por lo tanto, clave. La importancia de detectar estas afecciones en estas etapas iniciales ha determinado la orientación de la investigación en este ámbito y ha dado lugar a un buen número de artículos científicos sobre ENM pediátricas.

Esta producción textual es de especial interés para la investigación basada en corpus tanto por su carácter multidisciplinar, ya que en él confluyen numerosas especialidades del ámbito médico y sanitario, como por la variedad, actualidad y rigor de dichos textos.

2.1 Criterios de diseño

ENEUPECOR es un corpus bilingüe (francés-español), comparable, constituido por un subcorpus en cada lengua, y monogenérico, ya que está compuesto por artículos científicos.

Para su compilación hemos seguido los criterios que habitualmente se suelen tomar como base (Biber, 1993; Bowker, 1996; Meyer & Mackintosh, 1996; Sinclair, 1996; Pearson, 1998), si bien adaptándolos y primando algunos de ellos para adecuar la selección textual a nuestros intereses investigadores.

En esta ocasión la finalidad de nuestro estudio nos ha llevado a seleccionar textos redactados originalmente en francés o español. Este criterio que nos impusimos como restricción hace

que el corpus no incluya traducciones. De este modo se pueden analizar las características de textos producidos en situaciones comunicativas similares sin las posibles distorsiones originadas por las traducciones de un corpus paralelo.

En este punto debemos indicar que uno de los problemas para la creación del corpus ha sido que la producción textual original en francés sobre ENM es mucho más amplia que en español. Buena parte de los volúmenes de neuropediatría o de neurología pediátrica disponibles en castellano son traducciones del inglés. Sin embargo, a pesar de estas dificultades iniciales, conseguimos compilar una muestra textual representativa y de tamaño similar en ambas lenguas.

Para asegurar la comparabilidad y equilibrio la selección textual se ha llevado a cabo utilizando los mismos criterios en los dos subcorpus que conforman ENEUPECOR: número de muestras, temas abordados, cronología y extensión.

Los textos proceden de revistas especializadas de referencia, como *Archives de pédiatrie* (JCR, FI: 0.298) y *Anales de Pediatría* (JCR, FI: 0.770), lo que apoya el criterio de calidad. Son accesibles online y desde el punto de vista cronológico, si bien se incluyen en una franja temporal de 14 años, el 80% se sitúa entre los años 2000-2009. Se ha aplicado el criterio de representatividad en relación con la actualidad científica, aunque las características del subdominio temático nos han llevado a ampliar la cronología para incluir algunos textos de especial interés.

Los artículos seleccionados se caracterizan por tratar los temas abordados en profundidad. Entre otros aspectos, estos artículos hacen referencia a la evolución y desarrollo de las investigaciones en ENM, la aparición de nuevos criterios de clasificación, los resultados de nuevas pruebas de laboratorio que permiten diagnósticos más fiables y precisos, o la aplicación de posibles tratamientos.

2.2 Datos estadísticos del corpus

De acuerdo con los criterios indicados hemos elaborado un corpus cuyas características generales son las siguientes:

CRITERIOS	DESCRIPCIÓN DEL CORPUS
Canal	Textos escritos (formato electrónico)
Número de palabras	*Tokens*: 72 316 (fr) y 102 312 (es) *Types*: 9228 (fr) y 10 087 (es)
Contenido	Especializado tanto por el tema (textos con contenido específico ENM en pediatría) como por el tipo de género concreto (artículos científicos)
Tamaño de las muestras	Textos íntegros
Anotación	Corpus no anotado, muestras analizadas en formato .txt.
Límites cronológicos	Entre 1995 y 2009
Finalidad	Fin específico, recopilación de textos con fines léxicos y terminológicos
Lenguas	Bilingüe (francés y español)

Tabla 1. Datos generales del corpus

Como se puede apreciar, el número total de palabras (174 628) se corresponde con el límite superior de un corpus de tamaño mediano-pequeño (Vargas, 2006). Este número de palabras se ha contabilizado excluyendo tanto los resúmenes en otras lenguas como las referencias bibliográficas que figuraban en los textos. El objetivo buscado era poder establecer de este modo una cierta homogeneidad en la extensión de los textos que forman parte del corpus.

Si bien el debate sobre el tamaño que debe tener un corpus no está cerrado, son numerosos los autores que consideran que un elevado número de textos no es forzosamente sinónimo de calidad y representatividad (Kennedy, 1988; Leech, 1991). De modo más específico, encontramos autores, como por ejemplo Wright & Budin (1997) o Kock (1997), que destacan la utilidad y representatividad de los córpora de en torno a 100 000 palabras en ámbitos especializados, dado que el vocabulario utilizado es más restringido en estos casos.

De acuerdo con estas afirmaciones, el tamaño de ENEUPECOR es, por lo tanto, suficiente para conferirle el carácter de equilibrio y representatividad del ámbito temático seleccionado y del género textual en el que se centra.

El corpus compilado está constituido por un subcorpus en francés y otro en español, cada uno de los cuales cuenta con 25 muestras textuales respectivamente.

Los datos estadísticos del subcorpus en francés se resumen como sigue:

Tokens	72 316
Types	9228
Ratio Type/Token	12,76
Ratio Type/Token	12,76
Media (Tokens)	2763,40
Desviación estándar (Tokens)	1885,36

Tabla 2. Datos estadísticos del subcorpus en francés

En relación con el número de palabras de cada muestra textual, el intervalo general del subcorpus se sitúa entre 1236 y 9892 palabras. El 72% de los textos presenta entre 1800 y 6000 palabras y son mayoría en este grupo los textos en torno a las 3000 palabras. De los textos restantes un 24% incluye menos de 1800 palabras y solo el 4% más de 6000.

Podemos representar el número de palabras /texto en el subcorpus en francés como sigue:

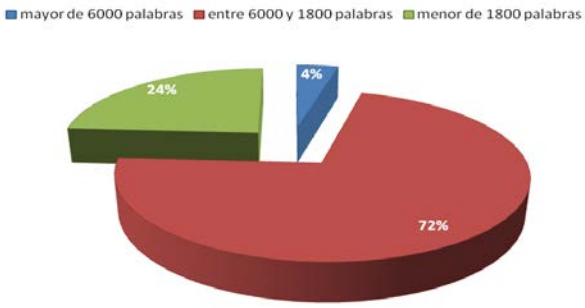

Figura 1. Intervalo número de palabras subcorpus en francés

Los datos estadísticos del subcorpus en español figuran en la siguiente tabla:

Tokens	102 312
Types	10 087
Ratio Type/Token	9,87
Estandarizado Type/Token	41,68
Media (Tokens)	4092,48
Desviación estándar (Tokens)	2029,43

Tabla 3. Datos estadísticos del subcorpus en español

En relación con el número de palabras de cada muestra textual, el intervalo se sitúa entre un mínimo de 1948 y un máximo de 10 160. El 88% de los textos presenta entre 1800 y 6000 palabras y solo el 12% supera esta cifra. Asimismo, la mayor parte de las muestras textuales de este subcorpus se sitúa en torno a los 4000 palabras (extensión media: 4092,48).

La distribución palabras/texto se presenta en el siguiente gráfico:

Figura 2. Intervalo número de palabras subcorpus en español

Como se puede observar, ambos subcorpus presentan un tamaño suficiente para asegurar su representatividad.

3 Metodología de explotación del corpus y resultados

Una vez creado el corpus procedimos a su explotación para elaborar un glosario terminológico bilingüe (francés-español) que incluyese los principales conceptos del subdominio de las ENM pediátricas.

En las páginas que siguen resumimos las cuestiones más relevantes en relación con la metodología seguida y las principales etapas del trabajo realizado. Estas etapas, si bien se presentan de manera secuencial, tienen generalmente, con excepción de la primera, un carácter recursivo.

3.1 Análisis de la formas más frecuentes

El primer paso en la explotación del corpus consistió en la extracción y análisis de las cincuenta formas de aparición más frecuentes. De este modo se pueden conocer los términos propios del subdominio objeto de estudio (Cabré, 1993; Dubuc, 1992; Rondeau, 1984) e identificar aquellas palabras que, por su frecuencia de aparición, podrían considerarse posibles candidatos a términos.

Para ello, utilizamos el conocido paquete informático *WordSmith Tools* diseñado por Michael Scott. Una vez realizada la exclusión de las palabras de contenido gramatical, con el fin de evitar el ruido de palabras vacías de contenido, obtuvimos dos listas, una por cada lengua de trabajo, que se recogen parcialmente en la tabla que sigue:

N	FR	Freq.	ES	Freq.
1	Maladie	270	Enfermedad	312
2	Musculaire	261	Muscular	534
4	Type	180	Tipo	129
6	Gène	175	Gen	123
7	Diagnostic	161	Diagnóstico	356
8	Clinique	151	Clínica	121
9	Patients	144	Pacientes	497
10	Ans	141	Años	181
11	Musculaires	139	Musculares	221
12	Myopathie	128	Miopatía	125
13	Maladies	125	Enfermedades	206
14	Enfants	123	Niños	122
15	Troubles	119	Trastornos	109
17	Étude	114	Estudio	251
18	Signes[2]	113		
19	Forme	112	Forma	153
21	Génétique	101	Genético	83
22	Mois	101	Mes	126
23	Âge	97	Edad	131
25	Muscles	94	Músculos	93
27	Syndrome	83	Síndrome	182
31	Tableau	81	Tabla	88
32	Dystrophie	79	Distrofia	221
34	Biopsie	73	Biopsia	169
36	Analyse	72	Análisis	83
39	Hypotonie	65	Hipotonía	292
40	Respiratoire	65	Respiratoria	112
46	Neuromusculaires	62	Neuromusculares	99
50	Duchenne	59	Duchenne	85

Tabla 4. Palabras de mayor frecuencia de los subcorpus en francés y español. Datos extraídos con *Wordlist*

Los resultados de este primer análisis nos permitieron, tras llevar a cabo una selección, obtener una relación inicial de posibles candidatos a términos monoléxicos. Como veremos más adelante (apartado 3.4) tomamos estas listas como punto de partida para realizar búsquedas más precisas y llevar a cabo el análisis exploratorio de cada una de estas formas en

[2] La celda en español está vacía porque la forma 'signos', equivalente de *sign*es, es la única de todas las recogidas que no se encuentra entre las 50 palabras más frecuentes también en el subcorpus en español.

su respectivo contexto de aparición. Esto nos permitió observar cómo funcionan los posibles candidatos en el interior del discurso, sus combinaciones y derivados.

La obtención de estas dos primeras listas muestra, asimismo, la existencia de un importante grado de coincidencia entre las formas más frecuentes en ambos subcorpus. De este modo se pone de manifiesto la homogeneidad de los mismos.

3.2 Extracción de contextos definitorios

Los contextos proporcionan pistas para saber cuándo estamos ante una unidad terminológica, dado que, en determinados casos, dichos contextos aportan información conceptual mediante la incorporación de definiciones.

El proceso seguido en esta fase fue diferente para el subcorpus en español y para el subcorpus en francés por la no disponibilidad de las mismas herramientas en ambas lenguas.

3.2.1 Extracción de contextos definitorios en el subcorpus en español con ECODE

Para extraer los contextos definitorios (CDs) en español hemos utilizado una herramienta de extracción automática del grupo de investigación del área de Ingeniería Lingüística de la Universidad Nacional Autónoma de México (UNAM), encabezado por Gerardo Sierra y Rodrigo Alarcón (Sierra *et al*., 2003 y 2008; Sierra, 2009). Estos investigadores han desarrollado un extractor automático de CDs basado en reglas lingüísticas (ECODE) y el sistema Describe® para la búsqueda, clasificación y agrupamiento de definiciones en la web.

La extracción de candidatos a términos con el ECODE necesita una gramática de patrones verbales definitorios (PVD) que contiene una serie de parámetros: verbos definitorios y los nexos que los acompañan; restricciones verbales referentes al tiempo y a la persona gramatical; patrones contextuales y restricciones de distancia entre el verbo y su nexo (Sierra, 2009).

Tras la extracción de los candidatos a partir de los PVD y el empleo de la gramática el análisis de los CDs requiere dos procesos principales: el primero consiste en eliminar los contextos irrelevantes mediante reglas de filtrado y el segundo en el reconocimiento de sus componentes.

Por último, el ECODE evalúa los CDs resultantes después de la etapa de filtrado, y en concreto los elementos constitutivos, para ponderar los mejores CDs según la estructura del contexto recuperado automáticamente. Se utilizan reglas heurísticas que contrastan las estructuras sintácticas de los elementos etiquetados como término y definición con sus estructuras prototípicas. Se designa un valor a cada elemento y un valor global a partir de las combinaciones localizadas. Los contextos que sobrepasen un umbral determinado serán los que el ECODE considere como buenos CDs.

Si bien el ECODE aún no está disponible en línea, tuvimos la oportunidad de que nuestro corpus en español fuese analizado con esta herramienta gracias a la disponibilidad, que agradecemos, de los investigadores de la UAM. Nos facilitaron información sobre las bases teóricas y el funcionamiento de su programa y trataron los textos. Los contextos obtenidos del análisis de ENEUPECOR, que servirán para aumentar el bagaje de experimentación de ECODE, nos permitieron obtener informaciones conceptuales importantes para localizar

posibles candidatos a términos, redactar las definiciones del glosario que queríamos elaborar y diseñar el árbol conceptual de nuestro subdominio.

A continuación presentamos algunos de los contextos definitorios en español obtenidos al final del tratamiento realizado. Hemos limpiado las etiquetas del ECODE, pero en cursiva se destacan los verbos y nexos que permitieron identificar los contextos definitorios (CDs):

Los trastornos neuromusculares *constituyen* un grupo de enfermedades que afectan cualquiera de los componentes de la unidad motora, *es decir*, la unidad funcional constituida por el cuerpo de la motoneurona del asta anterior de la médula espinal, su axón (nervio periférico) y todas las fibras musculares inervadas por esta motoneurona.
Las enfermedades neuromusculares hereditarias *son* trastornos heterogéneos en edad de inicio, clínica y gravedad.
Las distrofinopatías *son* trastornos causados por una anormalidad en el gen que codifica la proteína muscular distrofina.
La distrofia miotónica de Steinert *se caracteriza por* miotonía (dificultad o retardo en la relajación del músculo luego de su contracción), y debilidad muscular generalizada muy predominantemente facial.
La hipotonía neonatal generalizada *se define como* la disminución patológica del tono postural en las cuatro extremidades, el tronco y el cuello durante el primer mes de vida extrauterina
El síndrome de Walker-Warburg *consiste en* distrofia muscular congénita con anormalidades de las circunvoluciones cerebrales y cerebelosas, hipodensidad de la sustancia blanca en la TAC y alteraciones oculares (glaucoma congénito, hipoplasia retiniana y del nervio óptico, y cataratas).
Como amiotrofias musculares proximales *se conoce* a un grupo de enfermedades neuromusculares hereditarias de transmisión autosómica recesiva, ocasionadas por una degeneración de las motoneuronas del asta anterior de la médula espinal y posteriormente de los núcleos motores bulbares.
La terapia transgénica alternativa *utiliza* otros genes *para* restablecer la proteína deficitaria.

Tabla 5. Contextos definitorios en español extraídos por el ECODE

3.2.2 Extracción de contextos definitorios en el subcorpus en francés

Para la extracción de los contextos definitorios de los textos en francés no pudimos contar con el auxilio del ECODE, ya que esta herramienta, como indicamos, solo está disponible para español. Combinando el uso de *WordSmith Tools* con los fundamentos teóricos sobre los CDs, seleccionamos fragmentos de los textos en los que se nos daba información conceptual útil para también ahora identificar posibles candidatos a términos, redactar las definiciones del glosario en francés y realizar la estructuración del árbol en francés. He aquí algunos de los ejemplos más significativos:

La mutation du T521 *correspond à* une délétion d'une thymidine entraînant un décalage du cadre de lecture dans la partie du gène codant pour la portion extracellulaire de la protéine.
Le γ-sarcoglycane *est* une des protéines constituant le complexe de la dystrophine [5].
La thérapie génique *consiste à* apporter un vecteur contenant le gène d'intérêt et a pour objectif de corriger le déficit génétique par transfert du gène d'intérêt.
Les dystrophies musculaires et les myopathies métaboliques *représentent* les principaux groupes étiologiques.
La dermatomyosite juvénile *se manifeste par* un déficit musculaire et par des signes cutanés caractéristiques
Le terme « phénotype comportemental », qui *désigne* l'ensemble des comportements, mais aussi des troubles émotionnels et affectifs d'un sujet, peut paraître flou ou réducteur et ne doit être confondu avec « symptôme comportemental », il n'en reste pas moins le terme consacré en clinique
Les dystrophies musculaires des ceintures (LGMD) *forment* un groupe hétérogène de maladies, d'origine génétique, de gravité et de transmission variables et pouvant être divisé en deux groupes selon le mode d'hérédité : autosomique dominant (LGMD1, A-E) ou autosomique récessif (LGMD2, A-I).
Le complexe sarcoglycane *constitue* un premier sous-ensemble constitué de quatre protéines alpha, bêta, gamma, delta

Tabla 6. Contextos definitorios en francés

3.3. Elaboración de una organización conceptual bilingüe sobre ENM pediátricas

La organización conceptual es el esqueleto sobre el que se redactan las definiciones y que refleja el modo en que se estructura el área del conocimiento. Se trata de una de las fases más importantes, que, sin embargo, no ha estado exenta de problemas. Hemos tenido que plantearnos y tomar decisiones sobre cuestiones de diverso tipo para elaborar el árbol conceptual.

Por un lado el subdominio de las ENM pediátricas está en continuo cambio como resultado de las investigaciones y de los avances en la genética. Así, por ejemplo, hace años la clasificación de estas enfermedades se establecía a partir de la topografía, es decir, del componente de la unidad motora primariamente comprometido (enfermedades de la motoneurona, enfermedades del nervio periférico o neuropatías, enfermedades del músculo o miopatías y enfermedades de la unión neuromuscular). Sin embargo, en la actualidad se prefiere una clasificación basada en la biología molecular, que ha permitido la creación de nuevos subtipos dentro de un mismo conjunto de síntomas. Por este motivo hemos seguido esta última orientación para la clasificación de las enfermedades. Somos conscientes de que esta clasificación es susceptible de cambios en un corto período de tiempo, conforme vayan saliendo a la luz los nuevos hallazgos sobre las causas de cada una de ellas.

Por otro lado nuestra clasificación se ha visto condicionada por el hecho de que debería incluir únicamente las enfermedades que se manifiestan en la edad pediátrica. Por lo tanto enfermedades como la *polimiositis*, *miositis por cuerpos de inclusión* o *distrofia muscular oculofaríngea* se excluyen, ya que éstas solo se manifiestan en la edad adulta.

Asimismo, nuestro objetivo era incluir no solo la denominación de las enfermedades, sino también los conceptos relacionados con la patología, diagnóstico y tratamiento. Por lo tanto, el árbol conceptual tenía que abarcar todas estas áreas, además de la nosología (clasificación de las enfermedades), como se recoge en el siguiente esquema:

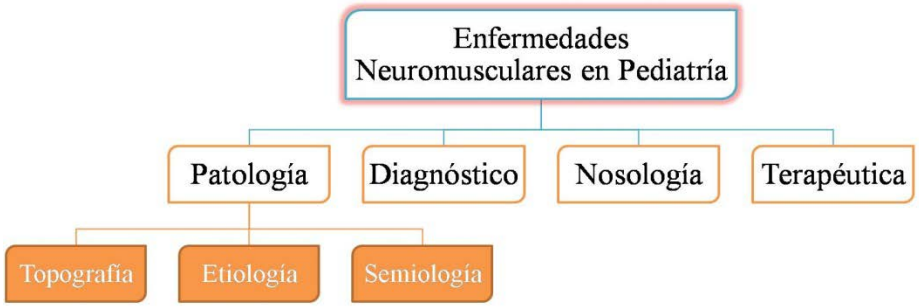

Figura 3. Esquema de las áreas de conocimiento incluidas

La elaboración del árbol conceptual conllevó un amplio proceso de documentación con fuentes especializadas (informes, monografías, artículos) y especialistas, que nos permitió conocer más a fondo el subdominio de las ENM pediátricas. El carácter multidisciplinar de este ámbito nos obligó además a trabajar con conceptos que estas afectaciones comparten con otras áreas como la genética o la anatomía.

A partir de las cuestiones indicadas, las lecturas de referencia y del análisis automático de los textos del corpus, validado en última instancia por los expertos en el tema, hemos elaborado el árbol conceptual y una estructura jerárquica bilingüe:

Enfermedades Neuromusculares en Pediatría	Maladies Neuromusculaires en Pédiatrie
1. Patología	1. Pathologie
1.1 Topografía	1.1 Topographie
1.1.1 Neuropatías	1.1.1 Neuropathies
1.1.1.1 Nervio periférico	1.1.1.1 Nerf périphérique
1.1.2 Enfermedades de la motoneurona	1.1.2 Maladies du motoneurone
1.1.2.1 Motoneurona	1.1.2.1 Motoneurone
1.1.3 Enfermedades de la unión neuromuscular	1.1.3 Maladies de la jonction neuromusculaire
1.1.3.1 Unión neuromuscular	1.1.3.1 Jonction neuromusculaire
1.1.4 Miopatías	1.1.4 Myopathies
1.1.4.1 Músculo	1.1.4.1 Muscle
1.2 Etiología	1.2 Étiologie
1.2.1 Enfermedades neuromusculares hereditarias	1.2.1 Maladies neuromusculaires héréditaires

Figura 4. Muestra estructura jerárquica bilingüe

3.4 Establecimiento de las denominaciones del glosario

Al empezar a trabajar más específicamente sobre los términos que íbamos a incluir en el glosario nos encontramos con dos aspectos que requirieron un tratamiento especial: las unidades sintagmáticas y la variación.

3.4.1 Identificación de las unidades sintagmáticas

La delimitación de lo que constituye una unidad terminológica, sobre todo en aquellos casos en los que especialistas han creado estructuras sintagmáticas para dar nombre a un concepto, no es fácil.

En primer lugar realizamos con la herramienta *Concord* de *WordSmith* una extracción de los *clusters* y los analizamos siguiendo las propuestas de Cabré (1993: 302). Esta autora ofrece una serie de pruebas que pueden ayudar a decidir si un segmento léxico corresponde a un término o si se trata de una combinación de términos.

Para ello tomamos como base el listado de posibles candidatos a términos simples que habíamos realizado (supra apartado 3.1), ya que el análisis de los *clusters* se lleva a cabo a

partir de un fichero de concordancias de una determinada palabra. Tomando como referencia el horizonte colocacional fijado, obtuvimos secuencias de palabras repetidas con una determinada extensión, *clusters* de dos, tres o cuatro palabras. Estos son los *clusters* de tamaño 2, 3 y 4 más frecuentes en francés y en español:

N	Word – Francés	Freq.
1	Dystrophie musculaire	55
2	Biopsie musculaire	48
3	Gène SMN	47
4	Maladies neuromusculaires	45

N	Word- Español	Freq.
1	Distrofia muscular	132
2	Debilidad muscular	83
3	Distrofias musculares	79
4	Biopsia muscular	74

Tabla 7. Los 4 *clusters* que más se repiten de tamaño 2

N	Word – Francés	Freq.
1	Prise en charge	55
2	Amyotrophie Spinale Infantile	23
3	Système nerveux central	21
4	Du système nerveux	20

N	Word- Español	Freq.
1	Muscular de Duchenne	56
2	Atrofia Muscular Espinal	30
3	Distrofia muscular congénita	26
4	Enfermedad de Werdnig	25

Tabla 8. Los 4 *clusters* que más se repiten de tamaño 3

N	Word – Francés	Freq.
1	Gène de la dystrophine	17
2	Dystrophie Myotonique de Steinert	14
3	Maladies neuromusculaires de l'enfant	14
4	ASI de type II	12

N	Word- Español	Freq.
1	Distrofia Muscular de Duchenne	55
2	Enfermedad de Werdnig-Hoffmann	22
3	Síndrome de Walker-Warburg	19
4	Mayoría de los casos	16

Tabla 9. Los 4 *clusters* que más se repiten de tamaño 4

Junto con este primer análisis aplicamos a ENEUPECOR las propuestas de Montero & Faber (2008: 81-87), que proponen una clasificación de las unidades de significación especializada (USE) atendiendo a diversos criterios y establecen una tipología de términos compuestos o unidades sintagmáticas presentes en los discursos especializados.

Así, en el corpus identificamos unidades lexicalizadas del dominio especializado (*distrofia facioescapulohumeral, Western blot,…*) y unidades fraseológicas usadas para parafrasear términos especializados o para cambiar de un registro formal a uno informal (*dolores musculares generalizados* y *mialgias, dolor de cabeza* y *cefalea,…*). Encontramos, asimismo, unidades fraseológicas de tipo metafórico, como las metáforas empleadas para caracterizar a las distintas afecciones o malformaciones de los pies (*pie de trinchera, pie de elefante, pie de atleta)* o las utilizadas para caracterizar los tipos de marcha (*marcha equina, marcha de pato, marcha laberíntica, marcha de gallo*). Obtuvimos también casos de unidades fraseológicas transmisoras de un solo concepto, equivalentes a términos compuestos, pero cuyos componentes carecen de fijación, admitiendo cambio de orden y variación, como *biopsia muscular* y *biopsia del músculo* o *cintura pelviana* y *cintura pélvica.*

Otro fenómeno analizado en esta fase fue el de la *truncación*, muy presente en sus diferentes posibilidades también en ENEUPECOR. Así identificamos la presencia de:

1. Siglas: *DMD* (*Distrofia muscular de Duchenne*), *DMB* (*Distrofia muscular de Becker*), *RM* (*resonancia magnética*), *MG* (*miastenia gravis*), *VCN* (*velocidad de conducción nerviosa*), *AME* (*atrofia muscular espinal*), *TC* o *TAC* (*tomografía computarizada* o *tomografía axial computarizada*),… Tanto en español como en francés se usan también habitualmente siglas en inglés, como *LGMD* (*Limb Girdle Muscular Dystrophy*).
2. Formas abreviadas: *quimio* (*quimioterapia*), *fisio* (*fisioterapia*), *maladie de Becker* (*dystrophie musculaire de Becker*),…
3. Acrónimos: *Agrimed* (*agricultura mediterránea*), *Insalud* (*Instituto Nacional de Salud*),…
4. Abreviaturas: *b.i.d.* (*bis in die* o *dos veces al día*), *q.h.* (*quaque hora* o *cada hora*), s.op.s. (*si opus sit* o *si es necesario*),…

Las unidades polilexemáticas o sintagmáticas se han identificado en el glosario como sintagmas nominales (S. nom.), ya que se trata de caracterizarlas desde el punto de vista gramatical. Muchas de estas estructuras se correspondían con denominaciones eponímicas de las enfermedades, como, por ejemplo, *distrofia muscular de Duchenne*, *distrofia muscular de Becker*, *enfermedades de Charcot-Marie-Tooth o distrofia muscular de Emery Dreifuss*. Nos encontramos también con casos de sustantivos seguidos de un adjetivo o de un sintagma preposicional que precisa su significado (*biopsia muscular*, *biopsia de nervio*). En algunos casos, cuando era pertinente, en la fase de elaboración del glosario se ha consignado una ficha a parte para el sustantivo que constituye el núcleo de ese sintagma y otra para el sintagma en cuestión (*miopatías* y *miopatías metabólicas*, *miopatías inflamatorias*,…).

Es interesante subrayar que, en nombre de la economía del lenguaje, se suele abreviar un determinado sintagma utilizando solamente su primer componente, como es el caso de *distrofia muscular*, *atrofia muscular* (sustituible por *amiotrofia* con la forma del griego *myós* o *mio-* referido al músculo), *hipotrofia muscular*, *debilidad muscular*, que se truncan y aparecen sin el modificador *muscular*, aunque se sobreentiende por el ámbito de conocimiento al que pertenecen los textos que estos hacen referencia a todo lo relacionado con los músculos y no con otros órganos.

Hemos encontrado algunos casos, los menos frecuentes, en los que la combinación léxica solo tiene sentido en sí misma y su significado no es igual a la suma de significados de sus partes (*Western Blot*).

3.4.2 Tratamiento de la variación

El fenómeno de variación y/o sinonimia en terminología ha sido estudiado a partir de diferentes perspectivas con criterios y parámetros distintos, lo que se traduce en una pluralidad de tipologías y en divergencias en su definición.

Para analizar la variación en ENEUPECOR aplicamos la clasificación de Faulstich (2002). Esto nos permitió identificar variantes lingüísticas de tipo fonológico y gráfico (*kinesiterapia* y *cinesiterapia*; *creatincinasa* y *creatinkinasa)*, morfológico (*cintura pelviana* y *cintura pélvica*) y léxico (*distrofia muscular de Duchenne* y *distrofia de Duchenne)*, variantes de

registro geográfico (*tomografía axial computada* y *tomografía axial computadorizada*), o bien variantes coocurrentes (*afección* y *enfermedad*; *sujeto* e *individuo) y v*ariantes competitivas (*LGMD* y *DMC*; *testing manuel musculaire* y *MMT).*

Junto a todos estos casos que nos han planteado problemas debemos destacar de modo especial algunos otros, como por ejemplo, la existencia de varias denominaciones para un mismo concepto. En el caso de la denominación de las enfermedades es frecuente el uso de una denominación eponímica y de otra no eponímica (*amiotrofia espinal tipo I* o *Werdnig-Hoffmann*), además del uso de siglas (*AME* o *amiotrofia espinal*).

Para resolver estas cuestiones tuvimos que determinar la pertinencia de esas designaciones para la elaboración del glosario, así como para decidir qué término usar como principal y cuáles como variantes.

El criterio seguido para determinar cuál era el término principal fue el de la frecuencia de aparición en las obras, así como la jerarquización propuesta por aquellos autores que mencionaban más de una denominación para un mismo concepto. Nuestras decisiones fueron revisadas por expertos.

4 Elaboración de un glosario en francés y español sobre ENM pediátricas

La información extraída de los diferentes análisis del corpus se organizó y sintetizó en fichas terminológicas que recogen los datos de nuestro interés para elaborar el glosario. El fichero terminológico está compuesto por un total de 246 fichas (123 en español y 123 en francés):

F-75	FR	25/04/2010	12/05/2010
Calpaïnopathie		n. f.	
Dystrophies des ceintures en formes récessives par déficit de calpaïne.		PETIOT, P. y URTIZBEREA, J. A. (2004): «Diagnostic des maladies musculaires», en *EMC – Neurologie*, 1 (2), pp. 137-155. Artículo disponible en: <http://www.em-consulte.com/article/25517/diagnostic-des-maladies-musculaires>. [Fecha de consulta: 29-04-2010]	
La forme la plus fréquente des ceintures correspond à l'entité clinique qui avait été décrite par Erb, et qui est appelée aujourd'hui «*calpaïnopathie*», en raison de l'identification d'un déficit d'une protéase de la famille des calpaïnes.		BEHIN, A., y PRADAT, P. F. (2002): *Neurologie*. Rueil-Malmaison: Groupe Liaisons.	

Figura 5. Muestra de ficha en francés

E-04	ES	26/04/2010	12/05/2010
Calpainopatía		n. f.	
Miopatía de cinturas debida a la alteración o ausencia de la calpaína muscular.		ERAZO-TORRICELLI, R. (2004): «Actualización en distrofias musculares», en *Revista de Neurología*, 39 (9), pp. 860-871. Artículo disponible en: <http://www.neurologia.com/pdf/Web/3909/r090860.pdf>. [Fecha de consulta: 25-04-2010] KLEINSTEUBER, K. y AVARIA, M. DE LOS ÁNGELES (2005): «Enfermedades Neuromusculares en Pediatría». [En línea]. *Revista Pediatría Electrónica*, 2 (1), pp. 52-61, <http://www.revistapediatria.cl/vol2num1/pdf/9_enfermedades_neuromusculares.pdf>. [Fecha de consulta: 25-04-2010]	
Las *calpainopatías* (LGMD2A) se deben a la alteración o a la ausencia de una enzima específica del músculo esquelético: la calpaína muscular (gen CAPN3 en el cromosoma 15). Una proteasa calcio dependiente.		ZARRANZ, J. J. (2003): *Neurología*. Madrid: Elsevier Science.	

Figura 6. Muestra de ficha en español

A partir de estas fichas hemos elaborado el glosario bilingüe que reúne los principales conceptos del subdominio de las ENM pediátricas cuya publicación estamos preparando:

D

dermatomyosite. *n. f.* maladie inflammatoire du muscle, de caractère auto-immun, qui se manifeste par une inflammation au niveau de petits vaisseaux dans les muscles (myosite) et de la peau (dermatite), entraînant des manifestations caractéristiques comme la faiblesse musculaire et la douleur, surtout au niveau des muscles situés au niveau des hanches et des épaules et des éruptions cutanées sur le visage, les paupières, les articulations des doigts, les genoux et les coudes.

dermatomyosites juvenile. *s. nom.* forme de dermatomyosites dont les signes commencent à l'âge de 16 ans.

DMB. voir dystrophie musculaire de Becker

DMC. voir Dystrophies musculaires congénitales

DMC avec déficit primaire en mérosine. *s. nom.* maladie génétique qui se manifeste par une atteinte musculaire s'accompagnant, au moins au début, d'une élévation des enzymes musculaires CPK ; on observe, par ailleurs, des anomalies de la substance blanche au niveau du cerveau, visibles dès l'âge de un an, sans retard mental, ni malformations oculaires associés.

DMD. voir dystrophie musculaire de Duchenne

DMJ. voir Dermatomyosites juvenile

Figura 7. Muestra del glosario en francés

G

gastrostomía. *n. f. i*ntervención quirúrgica que consiste en la creación de una abertura permanente que comunica el estómago con la pared abdominal a la que se recurre en caso de obstrucción de las vías digestivas superiores.
glucogenosis. *n. f.* miopatías metabólicas que pueden aparecer en cualquier edad y que cursan con fatiga muscular, calambres y dolores al realizar esfuerzo.
glucogenosis tipo II. *s. nom.* véase enfermedad de Pompe.
glucogenosis tipo V. *s. nom.* véase enfermedad de McArdle.

Figura 8. Muestra del glosario en español

Como se puede apreciar en estas figuras de muestra, en el glosario se han incluido las siglas presentes en el corpus. Esta inclusión se deriva de la importancia de estas formas en el ámbito de la biomedicina y que, por lo tanto, se manifiesta asimismo en el subdominio objeto de estudio.

En el glosario se ha recogido asimismo el fenómeno de variación, que abordamos en el apartado precedente. Así, en la entrada del término considerado principal se incluye la indicación *Var.*, enumerando a continuación los términos o variantes considerados secundarios. A su vez, en el apartado de remisiones de cada una de esas variantes secundarias, se indica *Véase*, para que se consulte el término principal.

5 Conclusiones

En este trabajo hemos presentado las fases seguidas para la elaboración del corpus bilingüe comparable ENEUPECOR y su posterior explotación con la finalidad de crear un glosario terminológico de utilidad para la traducción de textos del ámbito de la biomedicina en francés y español.

En las páginas precedentes hemos mostrado la utilidad del trabajo con corpus para la creación de recursos terminológicos. Presentamos cómo el corpus sobre ENM pediátricas que hemos creado —compuesto por artículos científicos, con un total de 174 628 palabras y organizado en dos subcorpus en función de las lenguas indicadas— nos ha permitido reconstruir el esqueleto conceptual de este subdominio en torno a cuatro conjuntos principales: patología, diagnóstico, nosología y terapéutica. Esta organización conceptual elaborada nos proporcionó una visión clara de las características del campo abordado. Por otra parte, la importante densidad de información conceptual y terminológica de los textos del corpus hizo posible el reconocimiento de las unidades especializadas complejas, de las variantes para un mismo concepto y de los criterios adoptados por los especialistas para la clasificación de las ENM.

Hemos mostrado asimismo la utilidad para la explotación del corpus de una metodología sistemática y de algunas herramientas informáticas que facilitan las diferentes fases de trabajo seguidas. En especial por su novedad destacamos el programa ECODE, que demostró ser una herramienta útil para el desarrollo de diccionarios especializados y glosarios. Pudimos comprobar que esta herramienta, con la que obtuvimos de forma automática la información necesaria para elaborar la definición de aproximadamente el 60% de las entradas en nuestro glosario en español, agiliza el proceso de elaboración de recursos terminológicos.

Para finalizar, queremos destacar que tanto el subdominio temático abordado (ENM pediátricas) como la selección de las lenguas confieren a este estudio otro rasgo de interés. Por un lado, nos hemos centrado en un ámbito temático novedoso, multidisciplinar y en el que existe una demanda social de difusión de la información. Como hemos indicado, las ER, en el que se integran ENM pediátricas, constituyen en estos momentos una prioridad en las políticas de salud pública. Por otra parte, frente al predominio del inglés en la comunicación científica, hemos identificado una producción textual de calidad tanto en francés como en español.

Este trabajo constituye una nueva aportación dentro de la línea de investigación que estamos desarrollando sobre traducción de textos médicos, en concreto ER y que toma como base la elaboración de corpus multilingües para la creación de recursos para traductores.

Como señala Malmkjaer (2003: 119), el uso de corpus ha cambiado definitivamente el paradigma de investigación en traducción. Ha marcado un hito solo comparable a la formulación hecha por Gideon Toury de las normas y la redefinición del concepto de equivalencia,

Agradecimientos – Agradecemos especialmente al Dr. Marcos Madruga Garrido (pediatra de la Unidad de Neuropediatría del Hospital Universitario Virgen del Rocío, Sevilla-España) la colaboración prestada en la revisión y validación de las cuestiones terminológicas y conceptuales del ámbito analizado.

6 Referencias

Baker, M. *et al.* (1993): *Text and Technology*. John Benjamins: Amsterdam/Philadelphia.

Behrman, R. E., Kliegman, R. M. & Jenson, H. B. (2004): *Tratado de Pediatría Nelso*n. Elsevier: Madrid.

Bernardini, S., Stewart, D. & Zanettin, F. (2003): Corpora in translator education: An introduction. In F. Zanettin, S. Bernardini & D. Stewart (eds.): *Corpora in translator education* (pp. 1-13). St. Jerome: Manchester.

Biber, D. (1993): Representativeness in corpus design. *Literary and Linguistic Computing*, vol. 8 (4): 243-257.

Bowker, L. (1996): Towards a Corpus-based approach to terminography. T*erminography*, vol. 3 (1): 27-52.

Bowker, L. & Pearson, J. (2002): *Working with Specialized Language: a practical guide to using corpora.* Routledge: London/ New York.

Cabré, M. T. (1993): *La Terminología: Teoría, metodología, aplicaciones*. Editorial Antártida: Barcelona.

Dubuc, R. (1992): *Manuel pratique de terminologie*. Linguatech Éditeur: Quebec.

Faulstich, E. (2002): Variação em terminologia. Aspectos de socioterminologia. In G. Guerrero Ramos & M. F. Pérez Lagos (coords.): *Panorama actual de la terminología* (pp. 65-91). Colmares: Granada.

Kennedy, G. (1988): *An Introduction to Corpus Linguistics*. Longman: London/ New York.

Kock, J. de (1997): Gramática y corpus: los pronombres demostrativos. *Revista de Filología Románica*, vol.12 (1): 291-298. http://revistas.ucm.es/index.php/RFRM/article/view/RFRM9797120291A .

Leech, G. (1991): The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (eds.): *English corpuslinguistics: Studies in honour of Jan Svartvik* (pp. 8-29). Longman: London.

Malmkjaer, K. (2003): On a pseudo-subversive use of corpora in translation training. In F. Zanettin, S. Bernardini & D. Stewart (eds): *Corpora in Translator Education* (pp. 119-134) St. Jerome: Manchester.

Meyer, I. & Mackintosh, K. (1996): The Corpus from a Terminographer's Viewpoint. *International Journal of Corpus Linguistics*, vol. 1 (2): 257-285.

Miquel, J. & Sánchez, E. (2010): The social model of translation and its application to health-specialised search engines on the Internet. An example: the ASEM neuromuscular disease search engine. *Meta*, vol. 55 (2): 374-386. http://www.erudit.org/revue/meta/2010/v55/n2/044246ar.pdf .

Montero, S. & Faber, P. (2008): *Terminología para traductores e intérpretes*. Tragacantos: Granada.

Pearson, J. (1998): *Terms in Context*. John Benjamins: Amsterdam/Filadelfia.

Pérez, C. (2002): *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento*. CSIC / Elies: Madrid. http://elies.rediris.es/elies_18 .

Rondeau, G. (1984) : *Introduction à la terminologie (2e édition).* Gaétan Morin: Québec.

Sánchez, E. (2008) : MYOCOR: création d'un corpus bilingue et d'un moteur de recherche sur les maladies neuromusculaires. In Translators Association of China (ed.):.*XVIII FIT World Congress Proceedings-Actes: XVIII Congrès mondial Fédération Internationale des Traducteurs* (pp. 350-360). Foreing Languages Press: Beijing.

Shreve, G. M. (2002): Knowing Translation: Cognitive and Experiential Aspects of Translation Expertise from the Perspective of Expertise Studies. In A. Riccardi (ed.): *Translation Studies. Perspectives on an Emerging Discipline* (pp. 150-171). University Press: Cambridge.

Sierra, G. (2009): Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática*, 2: 13-38. http://linguamatica.com/index.php/linguamatica/article/view/38/30 .

Sierra, G. *et al*. (2003): Definitional Contexts Extraction from Specialized Texts. In *PALC 2003 Proceedings: Language, Corpora and E-Learning* (pp. 21-31). Peter Lang Publish: Frankfurt.

―. (2008): Definitional Verbal Patterns for Semantic Relation Extraction. *Terminology*, vol. 14 (1): 74-98.

Sinclair, J. M. (1991): *Corpus, Concordance, Collocation.* Oxford University Press: Oxford.

―. (1996): *Preliminary recommendation on Corpus Typology*. EAG-TCWG-ATYP/P. EAGLES: Pisa.

Sirén, S. & Hakkrarainen, K. (2002): Expertise in Translation. *Across Languages and Cultures*, vol. 3 (1): 71-82.

Tagnin, S. E. (2005): *O jeito que a gente diz: expressões convencionais e idiomáticas.* Disal: São Paulo.

Varela, T. & Sánchez, E. (2012): EMCOR: a medical corpus for terminological purposes. *JoSTrans, The Journal of Specialised Translation*, vol. 18: 139-159. http://www.jostrans.org/issue18/art_varela.php .

Vargas, C. (2006): Diseño de un corpus especializado con fines terminográficos: el corpus de la piedra natural. *Debate terminológico*, vol. 2 (7). http://rua.ua.es/dspace/handle/10045/9426 .

Wright, S. E. Y Budin**,** G. **(**1997): *Handbook of Terminology Management Vol. 1.* John Benjamins: Amsterdam.