TERMINTEGRAL: Una plataforma para la construcción de bases terminológicas y ontologías

M. Teresa Cabré
Universitat Pompeu Fabra IULA, Barcelona

Linguistics engineering is frequently based on a descriptive analysis of terms and their combinations in order to design tools to detect units and structures that can be relevant for knowledge extraction. The integration of all the tools needed by terminologists in a platform enables the transformation of specialised texts and the construction and enrichment of ontologies. In this paper we present the design of TERMINTEGRAL, an open multimodular platform for terminology work. TERMINTEGRAL has been constructed by reusing tools previously developed by IULATERM, a research group of IULA at the Pompeu Fabra University (Barcelona). TERMINTEGRAL enables the construction of terminological glossaries and ontologies on the basis of automatic text processing.

1. Introducción

No cabe duda de que en el marco de la ingeniería lingüística la terminología es una materia de gran importancia, tanto en su vertiente descriptiva como aplicada.

Desde el punto de vista descriptivo cada vez es mayor el número de descripciones de los términos y sus combinaciones en los distintos ámbitos especializados. A través de estos trabajos podemos conocer las constantes estructurales y combinatorias que ofrecen los términos y la fraseología de los campos de especialidad. La ingeniería lingüística se ha basado a menudo en estas descripciones para diseñar herramientas de tratamiento automático de la información especializada que permitan detectar unidades y estructuras relevantes para la extracción de conocimiento.

Así, se han usado diccionarios para detectar la presencia o ausencia de una unidad específica en un tipo de texto, patrones sintácticos para identificar fragmentos de texto correspondientes a unidades terminológicas o fraseológicas, categorías gramaticales para descartar determinadas combinaciones que no conducen a unidades terminológicas, o patrones morfológicos para detectar unidades terminológicas morfológicamente construidas en ámbitos especializados que presentan una morfología sistemática en la formación de términos de algunas clases semánticas.

Muestra de algunas de estas herramientas son, en el marco de nuestro equipo de investigación[1], la herramienta *SEXTAN* (Vivaldi 2000; Bach & Vivaldi 2004) para la detección de neologismos formales en la prensa[2], el sistema YATE (Vivaldi 2001; 2003b) para la detección de unidades terminológicas pluriverbales, del tipo *sistema nervioso central, arteria aorta, bacilo de Kock* o *síndrome de Down*, en textos de medicina, o la utilización

de formantes grecolatinos en *YATE* (Estopà, Vivaldi & Cabré 2000a, 2000b) para la detección de unidades terminológicas univerbales compuestas, del tipo *cistitis, oligofrenia, corticoide, reumatología* o *cardiopatología,* en este mismo ámbito*)*, o *MERCEDES* (Vivaldi 2003a; Araya & Vivaldi 2004) que aparte de cooperar en la detección de unidades terminológicas, asiste al establecimiento de relaciones conceptuales, y a la construcción y actualización de los diccionarios de tratamiento de la información, ya sean generales o temáticamente especializados.

En su función de detección de unidades terminológicas, *MERCEDES* sirve especialmente para unidades univerbales no construidas. Esta herramienta se basa en la utilización de diccionarios de materias elaborados mediante el trabajo terminográfico llevado a cabo por nuestros doctorandos e investigadores, importados de Internet o construidos automáticamente a partir de la extracción de información relativa a un ámbito temático a partir de artículos de diccionarios generales y enciclopedias (en este último caso desarrollados en el formato *Mercedes)*.

La integración de estas herramientas en una misma plataforma, la plataforma *TERMINTEGRAL*, facilitaría el paso de los textos especializados a la construcción o enriquecimiento de productos terminológicos y ontologías.

2. El proceso de trabajo en terminología

La entrada de las tecnologías de la información y la comunicación como herramientas del trabajo terminológico ha hecho cambiar la metodología que se estaba utilizando ortodoxamente en la construcción de glosarios monolingües y multilingües. De entrada, la informatización del proceso de trabajo ha puesto en cuestión no únicamente la validez exclusiva del método onomasiológico, sino el valor universal de uno de los fundamentos que constituían la teoría dominante de la terminología: la preexistencia del concepto a la denominación. La Teoría terminológica tradicional de Wüster, denominada Teoría General de la Terminología (TGT), y el modelo expandido (que en Cabré 2003 se ha denominado *Teoría General Expandida de la Terminología* (*TGET*)), desarrollada por sus continuadores, mantienen que la delimitación del concepto es previa a la búsqueda de la unidad terminológica. Este fundamento sólo se sostiene en un proceso llevado a cabo por el especialista en la materia, el cual, al realizar un trabajo de recopilación terminológica, parte de sus conocimientos sobre la materia, conocimientos que posee estructurados en su mente. Sin embargo, un terminógrafo[3] con formación en lingüística o traducción partirá invariablemente del texto para iniciar su trabajo de detección y recopilación de unidades terminológicas, y aun para adquirir conocimiento sobre una materia. Este es el proceso que siguen también los estudiantes de cualquier especialidad que reciben lecciones de sus maestros o realizan lecturas de textos y gracias a los textos (orales o escritos) devienen poco a poco especialistas. Pero la utilización de herramientas que automati-

zan el proceso del trabajo terminológico convierte necesariamente la metodología terminográfica en semasiológica, ya que parten obligatoriamente del texto.

Se suele organizar el proceso tradicional de elaboración de una terminología en distintas fases[4]:

I. Fase preparatoria
II. Fase de elaboración de la terminología
III. Fase de revisión
IV. Fase de edición.

En la Fase preparatoria se distinguen distintas subfases:

1. La definición y delimitación del trabajo (ámbito temático, tema, destinatarios, funciones)
2. La búsqueda de información sobre la documentación disponible (bases de datos documentales, centros de gestión terminológica, bibliografías, etc.), sobre el tema, sobre los recursos terminológicos ya desarrollados y disponibles en todas las lenguas de trabajo (diccionarios generales y especializados, enciclopedias, normas, bases de datos, bases de conocimiento, etc.), y documentos sobre las características del ámbito profesional.
3. La organización de la información seleccionada: documentación sobre la materia para adquirir conocimiento sobre la misma y estructurar el ámbito de acuerdo con la delimitación del trabajo, documentación de consulta – monolingüe y multilingüe – sobre los términos de la especialidad, documentación para la extracción de términos – que compondrá el corpus de extracción o vaciado.
4. La organización del proceso de trabajo (constitución del equipo, distribución de funciones, fijación de la metodología, asesores, etc.)
5. Creación de los materiales de trabajo: diseño y organización de la base de datos, descripción de los campos de cada registro y redacción del protocolo de selección y representación de los términos seleccionados.
6. Redacción del plan de trabajo.

En la Fase II se elabora propiamente la base de datos terminológicos. Esta fase comprende:

1. La detección, en el corpus de extracción, de las unidades terminológicas pertinentes
2. La extracción de estas unidades y de la información adecuada sobre ellas (fuente, contexto(s), variantes, etc.)
3. La creación del registro para cada unidad terminológica seleccionada, siguiendo el protocolo de selección y representación elaborado
4. La complementación de la información de cada registro utilizando documentación de consulta.

En la Fase III de revisión del proyecto se corrigen los datos que corresponden a errores y se recuperan los registros problemáticos por distintos motivos:

1. por ser incompletos
2. por contener datos dudosos que hay que consultar
3. por no ser pertinentes desde el punto de vista del tema o de la selección de los términos.

Finalmente, una base de datos terminológicos puede dar lugar o bien a un diccionario, vocabulario o léxico, o bien a distintos vocabularios (cada uno producto de una selección de datos determinada por la delimitación de un tema, unos destinatarios y unas funciones), o bien servir de base de información para consultas en línea o en CD-ROM. En los dos primeros casos hay que decidir cómo presentar los datos: selección de los mismos, forma de representación, orden de las informaciones e inclusión de información complementaria, incluida la gráfica. En caso de tratarse de una base de datos consultable, hay que diseñar los tipos de consulta, las formas de acceso a cada tipo y la representación de la información que se obtiene en cada caso.

3. El proceso de trabajo automatizado

La diferencia metodológica que se da entre un trabajo realizado manualmente, aunque cuente con la informatización de los datos terminológicos en una base de datos, y un trabajo automatizado radica en principio en el hecho de que sea el ordenador y no el humano quien lleve a cabo cada una de las actividades del proceso. Para ello, el sistema debe disponer de instrucciones para realizar las distintas tareas del proceso y de la información adecuada y suficiente para poder actuar correcta y adecuadamente en cada caso. En suma, debe disponer de herramientas y de datos.

Revisamos a continuación las distintas actividades del proceso terminográfico a la vista de su posible automatización[5] y presentamos las necesidades del sistema para poder llevar a cabo directamente cada una de ellas.

Dentro de la I Fase, que consiste en la definición y preparación del trabajo, existen algunas tareas que sólo puede llevar a cabo un humano: estas tareas son tres: la definición y delimitación del trabajo, la organización del equipo y la redacción –aunque asistida– del plan de trabajo.

1. La primera tarea del proceso, que debe ser manual aunque puede contar con asistencia al acceso de trabajos parecidos que existen en red, consiste en la definición y delimitación del trabajo, y comprende:

- la elección del ámbito temático
- la delimitación del tema dentro de este ámbito

- la caracterización de los destinatarios tipo
- las funciones que el trabajo terminográfico concreto pretende cubrir en relación a necesidades de sus destinatarios.

Los parámetros seleccionados en esta tarea serán determinantes para la adecuada realización de las posteriores, algunas de las cuales pueden ser objeto de automatización.

2. La búsqueda de información puede automatizarse mediante el uso de un buscador temático, que explore vía Internet los centros documentales, los organismos especializados en el tema y las publicaciones monográficas y en serie, a fin de proporcionar al terminógrafo:

- Información sobre la documentación existente y disponible que por encima de todo sea pertinente para el trabajo diseñado (información documental), y sobre los organismos y entidades pertinentes para el tema (información factográfica).
  Para lograr una buena selección de referencias documentales, el motor de búsqueda de documentos debe poseer varios filtros: un primer filtro de pertinencia temática, un segundo filtro de adecuación de nivel de especialización y un tercer filtro de fiabilidad o calidad de los datos. Este motor de búsqueda facilita una lista de los documentos, ordenada de más a menos recomendable, que considera más adecuados para cada tipo de tarea a realizar. Para la selección de la información factográfica, el motor de búsqueda debe poseer también algunos criterios que le permitan ofrecer los datos ordenados por orden de pertinencia o calidad.
- Documentos sobre el tema y sobre la materia o materias en las que se suele encuadrar el tema objeto del trabajo. Entre estos documentos cabe distinguir los tipos siguientes, que no se excluyen necesariamente entre sí:

  - Fuentes especializadas: documentos que se utilizarán para adquirir información sobre el tema, que al mismo tiempo servirán también como fuente de consulta indirecta de los términos. Una selección de estos documentos constituirá el Corpus de extracción de las unidades terminológicas, los cuales proporcionarán al mismo tiempo los contextos de uso del término.
  - Fuentes de consulta directa: documentos que se utilizarán para resolver cuestiones sobre los términos: bases de datos terminológicos, diccionarios especializados en una materia o temática, diccionarios generales de ámbitos especializados, enciclopedias, normas, bancos de conocimiento sobre el tema y diccionarios generales de referencia
  - Fuentes factográficas: enlaces a centros de gestión terminológica que admitan consultas externas y a páginas de especialistas que puedan asesorar al terminógrafo.

3. La información seleccionada de cada tipo pasará a formar parte de una base documental y, en el caso de las fuentes especializadas pasarán a integrarse en una base textual. Los documentos integrados en la base de datos textual, que van a formar parte del corpus de extracción de los términos se procesarán automáticamente tanto estructural como lingüísticamente. El acceso a las fuentes de consulta directa (lexicográficas y terminológicas) y a las factográficas se hace mediante un enlace, a no ser que pueda disponerse de ellas en formato digital, como archivo residente integrado en la plataforma de trabajo del terminógrafo.
Las herramientas necesarias para llevar a cabo estas actividades son las siguientes:

- un clasificador de tipos de documentos que explore las fuentes seleccionadas y las asocie a un tipo de documento, siguiendo criterios de reconocimiento de tipos de documentos
- un conjunto de criterios de evaluación que ordenen dentro de cada tipo los distintos documentos por orden de pertinencia, adecuación y fiabilidad.

4. La organización del proceso de trabajo (constitución del equipo, distribución de funciones, fijación de la metodología, asesores, etc.) es el segundo bloque de tareas que debe realizar manualmente el terminógrafo, pero, una vez establecida esta organización, podrán automatizarse las vías de interacción entre el terminógrafo y cada uno de sus colaboradores según las funciones que ejerza cada cual. Para ello se crearán perfiles de colaborador y a cada perfil se le atribuirán unas posibilidades de acceso y/o modificación de la información constituida.
5. La creación de los materiales de trabajo, que se compone del diseño de la base de datos, la descripción de los campos de cada registro y la redacción del protocolo de selección y representación de los términos seleccionados, debe realizarse manualmente, aunque contando con un acceso fácil a las fuentes de información (enlaces a organismos que trabajen en terminología para poder consultar sus materiales de trabajo, acceso directo a las normas de terminología, publicaciones de recopilaciones de materiales terminográficos y acceso a bases de datos para analizar su estructura). Son especialmente relevantes en este caso las normas de terminología elaboradas por el CT/37 de ISO que ofrecen una lista muy exhaustiva de las categorías de datos terminológicos y formatos de representación y transferencia a nivel internacional.
6. Finalmente, la redacción del plan de trabajo, que cierra las tareas propias de esta I Fase de carácter preliminar, podrá automatizarse parcialmente mediante la utilización de una plantilla digitalizada correspondiente al tipo de documento *Plan de trabajo,* cuyos epígrafes se llenarán con los datos seleccionados en cada una de las tareas realizadas anteriormente, datos que, para optimizar su reutilización, podrían almacenarse en una base estructura en tipos de información requeridos por el formato del Plan de trabajo.

La Fase II del trabajo terminológico consiste en la elaboración propiamente dicha de la base de datos terminológicos y es en esta fase donde se han producido mayor número de herramientas de automatización del trabajo. Presentamos a continuación sus posibilidades a través de la lista de tareas descritas en el epígrafe anterior de este artículo. La detección de las unidades terminológicas en el corpus de trabajo puede llevarse a cabo mediante extractores automáticos de terminología, de los cuales existen bastantes en el mercado y en los ámbitos académicos. Estos extractores deben permitir reconocer:

- Las unidades terminológicas univerbales
- Las unidades terminológicas pluriverbales
- Las variantes formales de unas y otras.

El reconocimiento de las unidades univerbales (simples, derivadas y compuestas) es en general bastante complejo, por lo menos por dos motivos. El primero es porque no existe en la forma de estas unidades ningún indicio que permita reconocerlas como términos de un ámbito especializado, exceptuando algunas unidades derivadas y compuestas cultas que siguen una morfología recurrente en algún ámbito preciso, por ejemplo la medicina o la química. El segundo motivo es porque muchas de las unidades terminológicas coinciden formalmente con unidades léxicas de carácter general, ya que es su uso en un ámbito de especialidad el que determina su carácter terminológico. En consecuencia no existe en ellas elemento formal alguno que las identifique como términos, aunque mediante un analizador de contextos podrían arbitrarse elementos que las especificaran como términos. Para ello sería preciso trabajar sobre un corpus de textos etiquetados morfológica, sintáctica y semánticamente, lo que está todavía lejos de ser una realidad extendida.

La detección de unidades terminológicas univerbales de estructura morfológica simple (monomorfemáticas) suele llevarse a cabo mediante la incorporación de un listado de unidades temáticamente seleccionadas al detector, el cual compara si hay en el corpus textual cadenas de caracteres que coincidan con las unidades del listado y, si las hay, las identifica como términos. Estas unidades permiten a su vez recuperar unidades univerbales complejas (derivadas y compuestas) en las que aparece la unidad simple, ya que es muy frecuente en el léxico de los ámbitos encontrar series de unidades que por el hecho de compartir una misma raíz constituyen familias léxicas de carácter terminológico. Asimismo estas unidades simples, juntamente con las derivadas, permiten recuperar unidades sintagmáticas de carácter terminológico en las que aparece una unidad más simple de la misma familia, ya sea como núcleo del sintagma (en este caso serán casi automáticamente unidades terminológicas) o como complemento de un sintagma cuyo núcleo es una unidad semánticamente no especializada (con lo que seguramente detectaremos fraseología o colocaciones de un término dentro de un ámbito especializado).

Las unidades sintagmáticas o pluriverbales pueden detectarse también mediante la búsqueda automática de patrones estructurales con subespecificación semántica. Para ello hay que elaborar previamente patrones combinatorios de clases léxico-semánticas del ámbito de especialidad –del tipo *enfermedades, partes del cuerpo, sistemas funcionales en el organismo, localizaciones de los órganos*, etc.). Ello requiere disponer de un corpus etiquetado en cuanto a clases semánticas y de patrones de combinaciones semánticas recurrentes en cada ámbito de especialidad. Por ejemplo, en Medicina el patrón N*parte del cuerpo* + Adjetivo, tiene grandes posibilidades de ser un término; en cambio si N *dolor/enfermedad/lesión, etc*. + de + N*parte del cuerpo* probablemente no será una unidad terminológica sino, en todo caso, una unidad fraseológica propia de la medicina.

Al lado de la utilización de extractores que usan estrategias lingüísticas para la detección de términos, cada vez es mayor el número de adeptos a las estrategias estadísticas. Muestras de extractores sobre base estadística serían los sistemas ACABIT, WordSmith, TACT, NSP. La dificultad que puede impedir la construcción de un extractor de base estadística es que se necesita una enorme cantidad de datos de todos los ámbitos de especialidad para que el resultado sea mínimamente eficiente. Aparte de este obstáculo relativo a la creación de extractores, los existentes no están regularmente disponibles o bien no están entrenados para tratar datos de ámbitos especializados que no sean aquellos para los que se han construido. El sistema YATE (Vivaldi 2001) es un sistema mixto que utiliza pautas lingüísticas (estructurales y semánticas a través de WordNet) y pruebas estadísticas y que, aunque está adaptado para el ámbito de la Medicina, puede utilizarse en otros ámbitos siguiendo unas pautas de adaptación descritas explícitamente en el Manual de YATE (Vivaldi 2003b). Con todo, la totalidad de extractores de terminología proporcionan "candidatos a unidades terminológicas" que debe validar el terminógrafo o recurrir a la ayuda de los especialistas. Las unidades seleccionadas como candidatas a términos, una vez validadas por el experto o el terminólogo, pasan a constituir la entrada de un registro de la base de datos terminológicos.

Las bases de datos terminológicos estándar, sin embargo, no se limitan a una lista de unidades terminológicas que componen la entrada del registro, sino que llevan asociadas varias categorías de datos seleccionados en función de las finalidades para las que el banco de términos se crea[6]. En la mayoría de casos, el propio extractor ofrece la posibilidad de pasar automáticamente la información desde el texto a los registros del banco de datos, cada tipo de información en su campo correspondiente. Este proceso puede ser directo (el sistema consigna automáticamente la fecha de la entrada de datos, el nombre de su autor que se ha declarado previamente o la referencia de la fuente si previamente se ha declarado) o requerir alguna operación previa sobre el texto (por ejemplo, señalar el fragmento de texto que va a constituir un contexto, a no ser que el sistema tenga la instrucción de considerar el punto como indicador de cambio de fragmento) o sobre la transformación de los datos (una operación de lematización convertirá automáticamente la

unidad terminológica que aparece en el texto en su forma gramatical en su correspondiente forma lemática). El extractor está pues asociado a la base de datos mediante una *pasarela* que contiene todas las operaciones intermedias que aseguren que la representación de los datos en los registros siga las formas canónicas establecidas en el protocolo.

La introducción de *variantes* es otra de las operaciones que pueden automatizarse parcialmente aunque cada unidad requiere una validación previa a su consignación en la base de datos. La búsqueda de patrones de identificación de variantes por medio de marcadores como *denominado, que se denomina, que denominamos* X, *traducido en* Y *por* X, se formulan en un texto procesado lingüísticamente a través del lema cuando afecta la totalidad de sus formas gramaticales. Este tipo de búsquedas a través de fórmulas o de cadenas de caracteres permite recuperar del texto fragmentos en los que es probable que figuren variantes denominativas de un término. El uso de marcadores, en este caso marcadores metalingüísticos de tipo reformulativo o parafrástico, es un camino posiblemente relevante para la extracción de información. Estas estrategias, sin embargo, requieren un grado importante de refinamiento si se pretende obtener información realmente relevante. La complementación de estas búsquedas con otras herramientas permite refinar la búsqueda de variantes, como presentaremos en el próximo apartado.

La complementación de la información de cada registro puede realizarse utilizando documentación de consulta, cuyo acceso directo facilita una plataforma como TERMINTEGRAL, o bien a través de operaciones que forman parte de lo que se denomina *Text Mining*. La obtención de fórmulas definitorias, variantes, símbolos equivalentes, etc. forma parte de las técnicas de *Text Mining*. Muchos son los equipos que en estos momentos están trabajando en esta línea, como puede verse en las comunicaciones del último congreso de LREC celebrado en Lisboa en mayo de 2004 (LREC-2004).

La redacción de definiciones a partir de la información extraída del texto (fórmulas definitorias y contextos) y de los diccionarios, a los que una plataforma de trabajo da acceso directo, es una operación que puede semiautomatizarse mediante el uso de plantillas de definición, muy utilizadas por grandes editoriales en la confección de diccionarios generales y especializados. Esta actividad, sin embargo, no es tan simple como pudiera parecer, ya que previamente a la aplicación de una plantilla es necesario contar con una estructura adecuada a cada clase gramatical, y, dentro de ella, a cada clase semántica. Y para que las definiciones resultantes sean de calidad será preciso, en el caso de los diccionarios especializados, adaptar las plantillas a clases y subclases semánticas de cada ámbito o tema especializado. Las plantillas de definición no únicamente permiten controlar la sistematicidad de la redacción de las mismas sino que además facilitan el paso de la definición a la creación paralela de una ontología.

En la Fase de revisión del proyecto se corrigen los datos que corresponden a errores y se recuperan los registros problemáticos por distintos motivos: por ser incompletos, por contener datos dudosos que hay que con-

sultar, por no ser pertinentes desde el punto de vista del tema o de la selección de los términos. La gestión de los registros que hay que revisar se hace automáticamente a partir de la localización de campos vacíos (registros incompletos) o bien a través de una nota explícita indicativa. La localización de registros no pertinentes sólo puede llevarse a cabo manualmente, es decir, consultando a un experto o bien con posterioridad a la construcción de la ontología, cuando una unidad no consigue anclarse en ellas ninguna de las relaciones declaradas en dicha estructura ontológica.

Finalmente, si una base de datos terminológicos se ha creado para generar directamente un diccionario, ya sea en formato impreso o digital, ya sea como producto para ser consultado en línea a través de Internet, requiere únicamente un programa de exportación de datos a un sistema adecuado de representación y autoedición. Sin embargo, la gran mayoría de bases de datos son multifuncionales, y se almacena en ellas información procedente de distintos trabajos para que pueda servir posteriormente para generar diferentes productos. Por lo tanto, si tenemos en cuenta que a distintos perfiles de necesidades correspondería una selección diferente de datos, así como una adecuada representación de cada categoría de datos asociada a la unidad terminológica, un programa de edición de diccionarios a partir de una base de datos requiere disponer de módulos de edición adaptados a cada perfil, de forma que a cada módulo le correspondiera no únicamente una selección de categorías de datos sino además una representación adecuada de los mismos.

4. La reutilización de herramientas para la construcción de la Plataforma TERMINTEGRAL

En el ámbito de las estaciones de trabajo (*work stations),* el grupo IULATERM ha llevado a cabo el diseño de una estación para el trabajo terminográfico, TERMINTEGRAL, cuya concepción se basa en cuatro puntos:

a) la descripción previa de las fases, tareas y elementos que requiere un determinado perfil profesional
b) una descripción realista de las tareas que pueden automatizarse a gran escala y en un período restringido de tiempo
c) la reutilización de los recursos y herramientas desarrollados
d) una concepción modular y abierta de cada sistema de forma que este pueda enriquecerse progresivamente con la inserción de nuevos módulos.

Para el diseño y desarrollo de la plataforma TERMINTEGRAL[7], orientada al trabajo terminográfico (elaboración de obras terminológicas y construcción de una ontología sobre el tema) el grupo cuenta con las si-guientes herramientas, algunas en fase de desarrollo:

1. Para la búsqueda y selección de textos pertinentes: *BUSCATEXTOS* y *BUSCATEMAS*
2. Para la detección de unidades terminológicas: *YATE, MERCEDES* y *SEXTAN*
3. Para la detección de variantes léxicas*: YATE* y*, MERCEDES*
4. Para la detección y establecimiento de relaciones conceptuales (implementación en desarrollo): *MERCEDES*
5. Para la búsqueda de contextos y fórmulas definitorias: *BWANANET* y *MERCEDES*
6. Para la representación de las estructuras de conceptos a partir de los términos y relaciones detectadas: *RG3D*
7. Para la construcción de bases de datos terminológicos: *DÍGIT, MERCEDES* y *UPF_TERM*
8. Para la selección de casos problemáticos: *UPF_TERM*
9. Para la edición virtual de obras terminológicas: *UPF_TERM*
10. Para la construcción de ontologías: *ONTOTERM,* módulo ontológico de la base de conocimiento *GENOMA.*

Véase a continuación una breve descripción de las utilidades de cada herramienta:

BUSCATEXTOS es una herramienta en fase de construcción que busca referencias de documentos de tema preciso en Internet utilizando distintas estrategias: lingüísticas (palabras clave), textuales (ubicación de las palabras clave) y estadísticas (frecuencia absoluta y relativa y comparación de frecuencias: contraste entre la frecuencia de una unidad en un corpus general y en corpus de tema específico, contraste entre de la frecuencia en corpus de distinto tema específico, densidad terminológica de un documento, etc.). Propone una lista ordenada de documentos a partir de distintos índices: *pertinencia temática, nivel de especialización* y *fiabilidad* y organizados a través de la herramienta ***CLASIFICADOR*** (en fase de diseño)en tres tipos: textos especializados, diccionarios y datos factográficos. Cada tipo pasa a constituir una base documental diferenciada. Los documentos que se definen como textos especializados constituyen una base textual de la que se seleccionaran los documentos más adecuados para constituir el *corpus de extracción*, cuyos documentos serán procesados automáticamente y de ellos se extraerán los términos del trabajo. El resultado final es siempre validado manualmente. Ambas herramientas están siendo desarrolladas por Cabré, Bach y Vivaldi.

YATE (*Yet Another Term Extractor)* es un sistema de extracción de candidatos a término nominales desarrollado por Jorge Vivaldi (Vivaldi 2001), que fue construido para la extracción de términos de medicina en textos previamente procesados lingüísticamente. Permite obtener una lista ordenada de candidatos a términos mediante la combinación de los resultados obtenidos con diferentes métodos de extracción y utiliza información semántica en los procesos de extracción (EuroWordNet[8] (NOTA 8)). Para la detec-

ción de los términos utiliza estrategias heterogéneas (lingüísticas y estadísticas), cuyos resultados integra en un resultado único a partir del cálculo de terminologicidad (*termhood*: grado en que una unidad léxica representa un término pertinente al dominio considerado). Dispone de un Manual para adaptarlo a otros dominios temáticos.

MERCEDES (Vivaldi 2003a; Araya & Vivaldi 2004) es otro sistema de reconocimiento multilingüe de unidades terminológicas (UT) que permite recuperar términos en contexto y fragmentos de textos que contienen uno o más términos. Está compuesto por dos módulos: un programa de reconocimiento y un módulo de diccionario de términos. El diccionario se ha desarrollado a partir de un diccionario base y se enriquece con otros módulos que corresponden a glosarios electrónicos obtenidos en Internet o incorporados directamente de una base de datos o de documentos en formato de texto. Dispone de un módulo de gestión de diccionarios de referencia que permiten aplicar el programa a diferentes ámbitos y a cualquiera de las tres lenguas con las que trabaja actualmente (inglés, español y catalán). Este módulo permite la creación de diccionarios específicos a partir de dominios y subdominios identificados por su marca temática en una obra lexicográfica más extensa.

SEXTAN es un sistema de extracción automática de neología formal, que permite detectar automáticamente candidatos a neologismos a partir de texto procesado lingüísticamente. La estrategia de *SEXTAN* es la comparación de cadenas de caracteres y lemas con los lemas de un grupo de diccionarios predefinidos en cada ocasión.

BWANANET es la herramienta de interrogación general del Corpus Técnico del IULA (©IULA). Este corpus está etiquetado con un paquete de herramientas desarrolladas por el Institut für Maschinelle Sprachverarbeitung de la Universität Stuttgart (Corpus Workbench). El IULA ha desarrollado la interfaz *BWANANET* que permite la interrogación del CT-IULA vía Internet. Permite buscar unidades simples (frecuencias y concordancias) y conjuntos de unidades que cumplan determinadas condiciones (caracteres, categorías gramaticales, etc.) a través de tres tipos de búsqueda: la concordancia simple, la estándar y la compleja. La Concordancia estándar permite la búsqueda de hasta doce unidades diferentes. Las interrogaciones pueden hacerse sobre la forma, el lema y/o la categoría morfológica de forma combinada. La opción Concordancia compleja es la que ofrece más posibilidades ya que permite utilizar buena parte de la potencialidad del lenguaje de interrogación CQP. Esta opción, además de las búsquedas propias de la búsqueda mediante la concordancia estándar, permite interrogar sobre un número ilimitado de unidades, sobre todos los tipos de combinaciones de formas, lemas y/o categorías, y llevar a cabo cálculos de frecuencias sobre formas, lemas o categorías, etc

DÍGIT es un programa que efectúa la extracción automática de unidades terminológicas a partir de diccionarios y enciclopedias que contengan marcas temáticas. La información puede pasar a formato de base de datos terminológicos o al formato *MERCEDES*.

UPF_TERM es un servidor de bases de datos que da acceso a la terminología recopilada en el IULA. El protocolo de trabajo prevé la posibilidad de incorporar trabajos elaborados originariamente con gestores terminológicos diferentes (Multiterm y Termstar), o formatos tabulados simples (Excel, tablas de WinWord). Ofrece un formato de registro de los datos terminológicos muy amplio que puede adaptarse salvando algunas categorías de datos que son obligatorias. Ofrece un *Protocolo* o Manual de representación de los datos que puede consultarse en *http://upfterm.upf.edu:8080/protca.htm.*

ONTOTERM® es un sistema de gestión de terminología y ontologías desarrollado por A. Moreno de la Universidad de Málaga (Moreno 1999). Permite construir simultáneamente una base de datos terminológicos y asociar cada término a una ontología, que acepta una gran variedad de tipos de relaciones conceptuales. Los registros terminológicos ofrecen información sobre la ontología. El grupo de IULATERM ha utilizado ONTOTERM para la gestión ontológica y terminológica del Proyecto GENOMA.

Véase, para finalizar, un gráfico que sintetiza las distintas actividades a llevar a cabo en un proceso de trabajo (gráfico nº 1) y el proceso de trabajo con incorporación de las herramientas, las operaciones y los recursos integrados en TERMINTEGRAL (gráfico nº 2)[9].

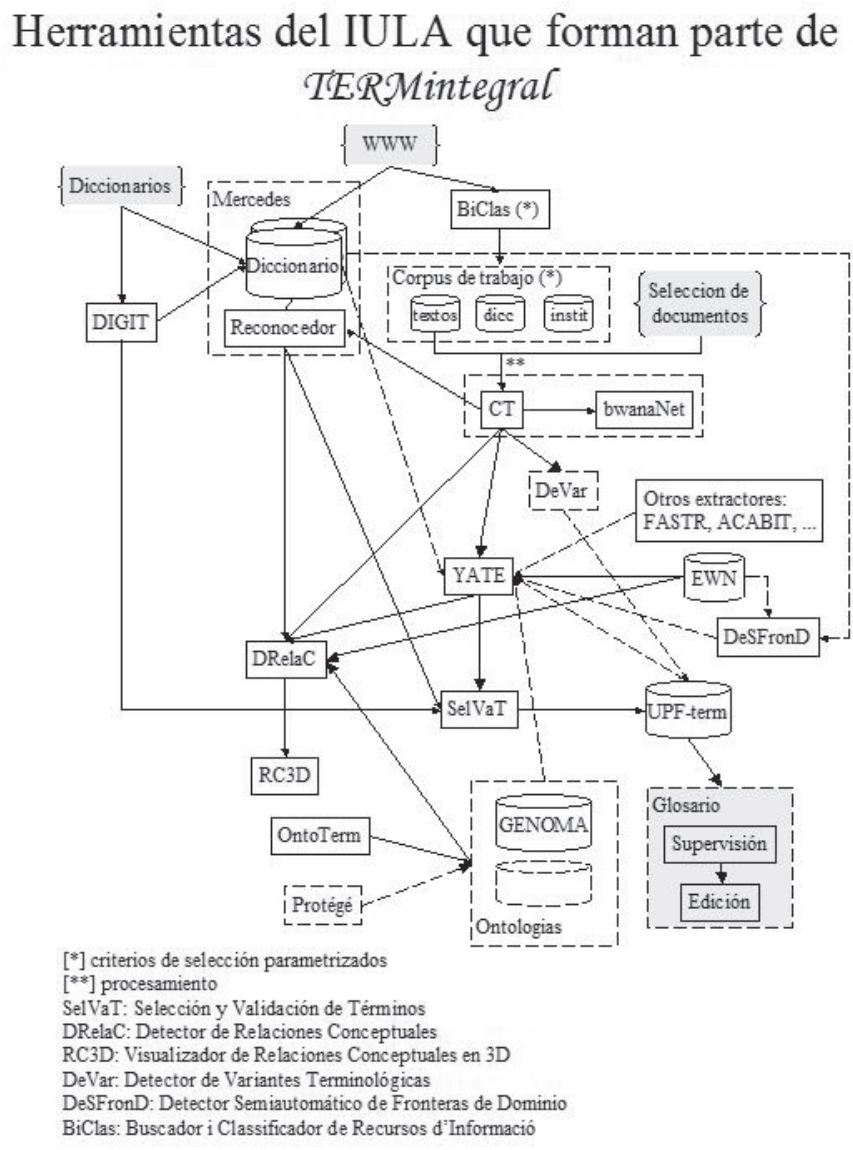

Gráfico 1: proceso de trabajo terminológico sistemático

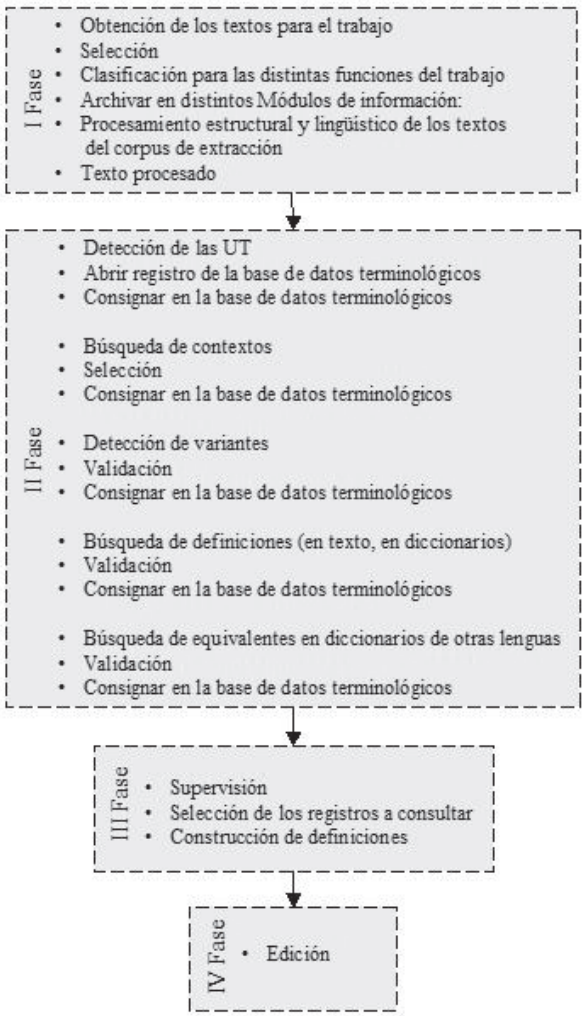

Gráfico 2:
*proceso de trabajo a través de la plataforma TERMINTEG*RAL

Bibliografía

Dadas las limitaciones de espacio y el carácter específico de este texto, sólo incluimos las publicaciones de miembros del grupo relacionadas con alguno de los programas o estrategias de TERMINTEGRAL, y las referencias citadas en el texto.

Araya, R. & J. Vivaldi (2004). "Mercedes: a term-context highlighter". *IV International Conference on Language Resources and Evaluation. LREC 2004*, 445-448.

Bach, C. & J. Vivaldi (2004). *Protocol per a la preparació de diaris de manera que es pugui treballar amb el programa de detecció de neologia SEXTAN*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [documento de trabajo interno]

Cabré, M. T (1999). "Informática y terminología". J. Mª. Blecua et al. (ed.): *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Editorial Milenio y Universitat Autònoma de Barcelona, 283-299.

Cabré, M. T. (2003). "Theories of terminology. Their description, prescription and explanation". *Terminology*, 9, 2, 163-200.

Estopà, R. (1999). *Extracció de Terminologia: Elements per a la Construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Tesis, 2). [Tesis doctoral publicada en CD-ROM, 2003]

Estopà, R. & J. Vivaldi M. T. Cabré (2000a). "Extraction of monolexical terminological units: requirement analysis". *Proceedings of the Workshop on Terminology Resources and Computation (LREC-2000)*. Atenas, 27 de mayo, 51-56.

Estopà, R. & J. Vivaldi & M. T. Cabré (2000b). "Use of Greek and Latin forms for term detection". *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC-2000)*. Atenas, 29 de mayo-1 de junio, 855-859.

Feliu, J. & M. Quixal (2002). *Manual d'ús d'OntoTerm. Versió 0.98.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [documento de trabajo interno]

Feliu, J. & M. T. Cabré (2002). "Conceptual relations in specialized texts: new typology and an extraction system proposal". *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, 45-49.

Feliu, J. & J. Vivaldi & M. T. Cabré (2002a). *Ontologies: a review*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (WP, 34). *ftp://ftp.iula.upf.es/pub/publicacions/02inf034.pdf* (consulted 20.08.2004)

Feliu, J. & J. Vivaldi & M. T. Cabré (2002b). "Towards an Ontology for a Human Genome Project". *LREC2002. Third International Conference on Language Resources and Evaluation. Proceedings*. Las Palmas de Gran Canaria, mayo de 2002, 1.885-1.890.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

LREC-2004 (2004). *Proceedings of the IV International Conference on Language Resources and Evaluation. LREC 2004.* Lisboa.

Moreno, A. (1999). "An introduction to OntoTerm". Workshop en el Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Barcelona, junio de 1999.

Vivaldi, J. (2000). "Sextan: prototip d'un sistema d'extracció de neologismes". Cabré, M. T. & J. Freixat & E. Solé (2000). (eds.) *La neologia en el tombant de segle. I Simposi sobre Neologia (18 de desembre de 1998) i I Seminari de Neologia (17 de febrer de 2000)*. Barcelona: Observatori de Neologia, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, 165-174. (Sèrie Activitats, 5).

Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Barcelona: Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. [Tesis doctoral]

Vivaldi, J. & H. Rodríguez (2002). "Medical Term Extraction using the EWN ontology". *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, 137-142.

Vivaldi, J. (2003a). *Sistema de reconocimiento de términos* Mercedes*. Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Vivaldi, J. (2003b). *Sistema de extracción de Candidatos a Término* YATE*. Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Vossen, P. (ed.) (1999). *EuroWordNet General Document*. University of Amsterdam. *http://www.hum.uva.nl/~ewn* (consulted 20.08.2004)

---

[1] El grupo de investigación IULATERM (Léxico, Terminología y Discurso especializado), del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra, se compone de investigadores y colaboradores de investigación. Para más información sobre el grupo y sus proyectos y actividades, véase http://www.iula.upf.edu/iulaterm.

[2] OBNEO es un banco de datos neológicos procedentes de un vaciado sistemático de prensa editada en español y en catalán desde el año 1989 hasta hoy. Puede consultarse en http://www.iula.upf.edu/obneo/obpreses.htm

[3] La vertiente aplicada de la Terminología, entendida como campo de conocimiento o disciplina que se ocupa de la descripción y explicación de las unidades terminológicas, se denomina Terminografía, utilizando el paralelismo con la Lexicografía en relación a la Lexicología. La denominación de terminógrafo o terminógrafa se aplica más genéricamente a quienes realizan recopilaciones de unidades terminológicas y las presentan en forma de diccionarios.

[4] La descripción de este proceso es una adaptación de la presentada en Cabré (1992).

[5] Sobre este tema puede consultarse además Cabré (1999).

[6] La creación de un banco de datos es un proceso costoso en tiempo y recursos, por lo que al diseñarlo se tiende a abrir al máximo sus posibilidades de utilización. Normalmente la adaptación de la información necesaria para cubrir perfiles de necesidades profesionales específicas se resuelve mediante la selección de categorías de datos que pueden constituir paquetes de información selectiva asociados a los distintos perfiles profesionales. Para más información sobre necesidades de tipos de información según perfiles profesionales puede consultarse Estopà (1999).

[7] En los apartados 2 y 3 se ha presentado una síntesis breve del estudio pormenorizado sobre el proceso de trabajo en terminografía y sus necesidades y posibilidades de automatización. TERMINTEGRAL, sobre la base de este proceso, selecciona las herramientas disponibles y propone las estrategias más viables y acotadas en un tiempo razonable, teniendo en cuenta que la plataforma puede ser incesantemente mejorada con nuevos módulos.

[8] EWN (Vossen P., 1999) es una base de datos léxica multilingüe de propósito general basada en el WN de Princeton (Fellbaum, 1998) que abarca tanto el español y el catalán como otras lenguas europeas. Cada idioma tiene su propia estructura *WordNet*, mientras que todas las lenguas incluidas en el proyecto están enlazadas entre sí por medio de una estructura común.

[9] Debo a J. Vivaldi la elaboración de los diagramas y a J. Feliu la revisión formal del texto y la bibliografía.