FunGramKB y la Adquisición Terminológica

Ángel Felices Lago, Universidad de Granada (España)
Pedro Ureña Gómez-Moreno, Universidad de Granada (España)
Ángela Alameda Hernández (España)

Índice

Resumen

Este artículo presenta de una manera práctica el proceso de recopilación y adquisición terminológica mediante la herramienta *FunGramKB Term Extractor*. Analizaremos paso a paso cómo, partiendo de un corpus de textos, obtenemos y definimos un conjunto de términos especializados representativos de un área temática concreta. La importancia de la adquisición terminológica es doble: por un lado, los términos especializados servirán como sustrato lingüístico, tanto para la definición de conceptos como para la creación de subontologías vinculadas a la Ontología Nuclear de la base de conocimiento FunGramKB; en segundo lugar, el trabajo terminológico permitirá no sólo el poblamiento conceptual de un dominio terminológico, sino también la creación de glosarios y diccionarios especializados para labores traductológicas o didácticas.

Palabras clave: FunGramKB, base de conocimiento, ontología satélite, procesamiento del lenguaje natural, extracción terminológica, lingüística de corpus

1 Introducción

En referencias precedentes ya se ha explicado ampliamente que FunGramKB es una base de conocimiento léxico-conceptual multipropósito diseñada especialmente para su uso en sistemas del PLN, fundamentalmente para tareas que requieran la comprensión del lenguaje (cf. Periñán-Pascual y Arcas-Túnez 2004; Periñán-Pascual y Arcas-Túnez 2005; Mairal-Usón y Periñán-Pascual 2009; Periñán-Pascual y Mairal-Usón 2009). En este estudio, partimos de la hipótesis de que el modelo multinivel de la

Ontología Nuclear de FunGramKB (i.e. nivel metaconceptual, nivel básico y nivel terminal) puede ser exportado a un modelo subontológico terminológico con el fin de minimizar la redundancia informativa y maximizar el conocimiento, tal y como ya ocurre en la Ontología Nuclear (Mairal-Usón, Periñán-Pascual y Samaniego-Fernández 2011; Felices-Lago y Marín-Rubiales [en prensa]). Implementando el mismo tipo de arquitectura y el mismo lenguaje de representación del conocimiento (i.e. COREL) en el nivel conceptual terminológico no sólo dotamos a FunGramKB de una mayor consistencia interna, sino que además posibilitamos la reutilización del mismo motor de razonamiento diseñado para la Ontología Nuclear, permitiendo su aplicación en tareas de comprensión del lenguaje natural. El repositorio final resultante sobre la temática que seleccionemos (derecho, medicina, ingeniería, etc.) podrá reutilizarse tanto por los humanos (a través de una interfaz a modo de diccionario) como por la máquina (a través de su futura aplicación a sistemas del procesamiento del lenguaje natural (PLN)). Para ello será necesario identificar previamente las palabras que pertenecen al dominio temático seleccionado, además de estructurarlas jerárquicamente a partir de la recopilación y gestión de una amplia base de datos textual y/o documental on-line procedente de fuentes solventes.[1]

Actualmente no existe un acuerdo general sobre qué es un término y cómo ha de definirse. De forma inicial, definiremos "término" como una unidad léxico-conceptual perteneciente al discurso especializado de un dominio de conocimiento concreto, como por ejemplo el de la medicina, el derecho o la tecnología, y que no es característico de la lengua general (cf. Cabré 1999 y Temmerman 2000, entre otros). Por ejemplo, la palabra "leucocito" es un término, puesto que su uso se restringe al habla profesional y académica del ámbito médico.[2] De igual forma, la palabra "adamelita" es un término perteneciente al dominio de la geología y su uso se circunscribe inequívocamente a este ámbito.[3] En consecuencia, se denomina "terminología" al conjunto formado por los términos propios de un ámbito de conocimiento especializado.

Desde la etapa inicial de recopilación de los textos de los que posteriormente se extraerá la terminología, hasta la selección y definición de los términos, es necesario seguir una metodología exhaustiva que garantice la extracción del mayor número posible de términos. Básicamente, el proceso terminológico consta de tres fases: (i) compilación del corpus o repositorio de textos, (ii) extracción y (iii) definición de términos. En los apartados que siguen explicaremos cada una de estas etapas del proceso de adquisición o "extracción" terminológica. Para ello, se aportarán ejemplos provenientes de una subontología sobre crimen organizado y terrorismo basada en la extracción terminológica de una amplia colección de textos denominada *Global Crime Term Corpus* (en adelante, GCTC). Previamente, no obstante, es necesario que definamos cuál es la función del extractor terminológico, e introduzcamos algunas nociones básicas sobre terminología.

[1] La trilogía clásica de los pasos fundamentales para el análisis conceptual en la gestión terminológica (Wright y Budin 1997: 106-07) consiste en los siguientes axiomas básicos: (i) preparación del listado de términos relevantes; (ii) verificación/confirmación de las decisiones que se toman con otros colegas y los expertos en la materia; y (iii) construcción de las definiciones.

[2] Término extraído del diccionario médico-biológico, histórico y etimológico Dicciomed.eusal.es (www.dicciomed.es).

[3] Término extraído del glosario de geología de la Real Academia de Ciencias Exactas, Físicas y Naturales (www.ugr.es/~agcasco/personal/rac_geologia/rac.htm).

2 Estructura y funcionamiento de FunGramKB Extractor

El extractor terminológico de FunGramKB se encuentra integrado en la aplicación web de FunGramKB.[4] Su función principal consiste en el almacenamiento y procesamiento de textos con el propósito de producir a partir de ellos una lista de términos especializados. La importancia del extractor reside en que es una herramienta que facilita la identificación de términos de manera eficiente. El proceso de adquisición terminológica en FunGramKB es semiautomático o asistido, ya que, si bien el extractor propone automáticamente una lista de palabras que son potenciales términos especializados, es el terminólogo el que deberá decidir cuáles de los términos seleccionados por el extractor son terminología realmente.

Existen en la actualidad distintas metodologías y marcos teóricos para la extracción (semi)automática de términos. El procesamiento de los textos y la extracción terminológica en FunGramKB tienen un enfoque estadístico. El extractor está basado en el cálculo de la denominada *term frequency–inverse document frequency* (*tf-idf*) que mide la importancia estadística o peso específico de cada término en un corpus dado. Este cálculo, que se expresa mediante un índice numérico, es el resultado de ponderar o "normalizar" la aparición de una palabra en un documento (*tf*) con la aparición de esta misma palabra en los distintos documentos de los que se compone el corpus (*idf*). El tf-idf prioriza aquellos términos que tienen una frecuencia absoluta elevada pero un *idf* bajo, mientras que relega o descarta aquellos términos que ocurren en muchos (o la mayoría) de los documentos del corpus. De esta forma, cuanto mayor sea el índice *tf-idf* de un término, mayor la probabilidad de que sea un término especializado, o, en términos estadísticos, podrá considerarse un término significativo. Este índice matemático será el que guiará al terminólogo en la elaboración de una lista de términos especializados.

El extractor consta de tres partes principales (Figura 1): (i) el indizador o "preprocessing (indexing)", (ii) el procesador estadístico o "processing (statistics)" y (iii) el extractor terminológico, incluido en la función "view". El indizador permite la carga estructurada de los textos recopilados para su posterior procesamiento; el procesador estadístico es el encargado de realizar el cálculo matemático de los índices *tf-idf*, así como de presentar una lista ordenada de términos potencialmente relevantes; finalmente, el extractor terminológico es la herramienta que permite trabajar con los términos, ya sea para descartarlos o definirlos una vez que se hayan considerado relevantes.

Figura 1. Pantalla principal del extractor de FunGramKB

Otras herramientas secundarias del extractor son el motor de búsqueda específica o "search", que facilita la localización de estructuras lingüísticas concretas en los textos, y la herramienta denominada "corpus", cuya finalidad es la visualización y el recuento de los textos que han sido cargados en el corpus, así como el número total de palabras que lo componen.

[4] www.fungramkb.com.

3 Metodología para la extracción terminológica

3.1 Compilación de un corpus especializado

Un "corpus" se define de forma general como un conjunto de textos escritos (o transcritos del lenguaje hablado) cuya estructura y formato permiten su procesamiento computerizado con el fin de estudiarlos en sus dimensiones lingüísticas y conceptuales.[5] Un aspecto fundamental en la extracción de términos pertenecientes a un dominio de conocimiento especializado consiste en la recopilación (o "compilación") de un corpus de textos que sea representativo de dicho dominio. El corpus servirá como base para obtener una lista completa de términos y de conceptos especializados. Existen básicamente dos tipos de corpus: los de propósito general y los corpora especializados. Los corpora generales representan el lenguaje general y están formados por textos pertenecientes a distintos géneros lingüísticos, tales como el literario, el periodístico o el habla coloquial. Los corpora especializados, por el contrario, representan un tipo de lenguaje específico sobre un área de conocimiento concreta, como la medicina o la tecnología. Si nuestra labor consiste en la población de una subontología dentro de FunGramKB, es necesario disponer de un corpus representativo del dominio sobre el que se está trabajando. A continuación explicaremos cuáles son los pasos para la compilación del GCTC, que es un corpus especializado del dominio jurídico, y, en concreto, del subdominio del terrorismo y crimen organizado, y que está formado por tres lenguas: inglés, español e italiano.

3.1.1 Selección de las fuentes

El primer paso para la compilación del corpus consiste en la búsqueda y selección de las fuentes de información, es decir, aquellos repositorios académicos o entidades profesionales que contengan documentos que puedan conformar un cuerpo textual especializado. La selección de las fuentes, así como la de los textos que éstas ofrecen, es de suma importancia en el proceso de extracción, puesto que de ella dependerá que el corpus resulte óptimo, tanto cualitativa como cuantitativamente. Por este motivo, las fuentes de las que se nutrirá el corpus deberán ser de reconocido prestigio dentro del ámbito científico-técnico en el que se encuadren. En el caso del GCTC se han seleccionado fuentes especializadas en materia de cooperación internacional contra el terrorismo y el crimen organizado, tales como la Unión Europea, el Consejo de Europa o la Corte Penal Internacional.

3.1.2 Recopilación y formateo

Una vez concluida la selección de las fuentes, el siguiente paso consistirá en recopilar y almacenar electrónicamente los textos que éstas ofrecen. A fin de que el extractor pueda procesar el corpus, es necesario que los documentos recopilados tengan un formato que sea legible para el extractor. Éste reconoce el denominado "texto sencillo", que es el formato de los archivos de extensión "txt". Con frecuencia, las fuentes ofrecen sus documentos en "pdf", "html" y otros formatos propietarios muy comunes como "doc(x)". Es necesario, por tanto, en los casos que sea preciso, reformatear los archivos a texto sencillo para obtener un archivo que no contenga imágenes o caracteres especiales que no pertenezcan al estándar ASCII. En

[5] La bibliografía referente a la lingüística de corpus es muy extensa. Para definiciones de corpus, léanse principalmente Kennedy (1998), Reppen (1998) y Meyer (2002). Para referencias sobre creación de corpora especializados, léanse Repen (2010) y Koestler (2010), y referencias mencionadas en estos.

lingüística computacional se conoce con el término "ruido" (*boilerplate*) a los caracteres o etiquetas de metadatos que no son relevantes y que pueden ocasionar problemas en una determinada tarea de procesamiento. Nuestra labor en este punto será tratar de minimizar la aparición de elementos que no sean pertinentes para la de extracción de términos especializados.

3.1.3 Edición de los textos

Uno de los aspectos fundamentales en el procesamiento terminológico consiste en la edición de los textos que componen el corpus, esto es, en la corrección de aquellos errores que hayan podido surgir como resultado del proceso de formateo de los documentos. La importancia de la edición de los textos radica en que, cuanto mayor sea la corrección del documento, mayor será la posibilidad de recuperar información relevante en etapas posteriores de la extracción terminológica. Existen un número indefinido de posibles errores ortográficos y tipográficos que pueden ocurrir durante el proceso de formateo a texto sencillo. A continuación se muestran algunas de las tareas de edición más frecuentes (i-ix).

(i) Rediseño de textos de dos columnas. Convertir textos de dos columnas a una sola;
(ii) Corrección de errores ortográficos. Con frecuencia surgidos con “f”, “fl” o “fi”;
(iii) Corrección de intercalados erróneos entre el cuerpo del texto y los encabezados, membretes o títulos;
(iv) Corrección de intercalados erróneos de pies de páginas con el cuerpo del texto;
(v) Corrección de sílabas o letras de palabras separadas por espacios (por ejemplo, “f i e l d”);
(vi) Corrección de sílabas o letras erróneamente separadas y unidas a otras palabras (por ejemplo, “crimi nalactivity” en vez de “criminal activity”;
(vii) Corrección de palabras divididas silábicamente mediante un guión al final de línea;
(viii) Eliminación de enlaces web o direcciones “url”;
(ix) Eliminación de direcciones de correo electrónico;

Para que se pueda llevar a término en el menor plazo de tiempo posible, la corrección de cada una de las potenciales amenazas ha de realizarse de forma manual mediante el trabajo coordinado de un grupo de compiladores. Serán estos los que tendrán que evaluar el tiempo que necesitarán para realizar las correcciones, así como la posibilidad de resolver algunas de las tareas de corrección mediante la aplicación, individual o por lotes, de órdenes computerizadas que reparen estos errores automáticamente de forma total o parcial. Tal es la función de algunos programas informáticos para el tratamiento de documentos, como por ejemplo *Notepad++*.[6]

3.1.4 Registro y creación de una base de datos

Paralelamente a la recopilación del corpus, los compiladores habrán de crear una base de datos que incluya un registro de toda la información pertinente para cada uno de los textos que se van a utilizar en el extractor. La base de datos informatizada contendrá información relativa al título de cada texto, su contenido, la fuente de la que se ha

[6] Notepad++ es una herramienta de código abierto y disponible gratuitamente en http://notepad-plus-plus.org/

extraído u otra información que los compiladores consideren relevante. La Figura 2 muestra la base de datos que se ha utilizado para el registro de los textos del GCTC.

Figura 2. Base de datos: Claves para la identificación de las fuentes

ID	Language	Brief description	Title	Topic	Type of document	Source
1	English	Fight against organised crime	EOAct (joint) law enforcement cooperation	Organised Crime	Joint Action	Eur-Lex
2	Spanish	Fight against organised crime	SOAct (joint) law enforcement cooperation	Organised Crime	Joint Action	Eur-Lex

El primer campo “ID” asigna un único código numérico a cada texto. El campo “Language” contiene información respecto al idioma en el que está escrito el texto. “Brief description” ofrece un título sobre el tema general del que trata el texto y “Title” ofrece un título que representa el tema concreto del documento (véase el siguiente apartado sobre “nominación de los textos”). El campo “Topic” registra el subdominio al que pertenece el texto; en el caso de GCTC, se distingue entre “Organised crime”, “Terrorism” o “Both”. Finalmente, el campo “Type of document” se ha diseñado para contener información sobre el tipo de texto y “Source” añade información adicional sobre la fuente original de la que se extrajo el documento. La información recogida en la base de datos tiene dos objetivos. En primer lugar, servirá como guía interna de organización para los compiladores del corpus. En segundo lugar, algunos de los datos que se registran serán necesarios durante la fase de carga en el extractor (véase apartado 3.2.1 abajo). Finalmente, la base de datos servirá como base cualitativa en etapas posteriores que requieran la justificación documental del corpus, así como la presentación de los resultados de la extracción terminológica y la creación de la subontología.

3.1.5 Nominación de los textos

Como paso previo a la carga del corpus en el extractor, es necesario tomar algunas decisiones respecto al nombre que se le asignará a cada uno de los textos recopilados para que los identifique de forma inequívoca. No existe un criterio común para nombrar los archivos de texto, antes bien, cada proyecto de creación de corpus sigue sus propias convenciones y toma sus propias decisiones sobre qué elementos han de componer los títulos de los archivos. En lo que respecta al GCTC se han seguido dos criterios: en primer lugar, el nombre de los archivos ha de ser lo suficientemente distintivo como para minimizar el riesgo de coincidencias con los nombres de otros archivos; en segundo lugar, el nombre que se asigne ha de coincidir siempre con el que se registra en la base de datos, de tal manera que si uno se altera el otro ha de cambiar de manera correspondiente. Finalmente, es necesario resaltar que, tanto la nomenclatura, como el registro de los textos en la base de datos, deben utilizar el inglés como lengua común, independientemente del idioma de los documentos o el componente del corpus que se esté compilando.

Para la compilación del GCTC se ha seguido el sistema de nomenclatura que aparece ilustrado en los ejemplos (1), (2) y (3).

(1) “ETDeci combating terrorism”
(2) “SORes persons traffic supression”
(3) “IBRep anti money laundering”

Donde:

(i) La primera letra indica la lengua en la que está escrito el texto. En este caso, “S” (*Spanish*), “E” (*English*) o “I” (*Italian*);

(ii) La segunda letra indica el subdominio al que pertenece el texto. En este caso, "T" (*terrorism*), "O" (*organized crime*) o "B" *(both*);

(iii) La tres últimas letras de la primera palabra se refieren al tipo de documento. En este caso, "Res" (*resolution*), "Deci" (*decision*) y "Rep" (*report*). En el primero de los casos aparecen cuatro letras en lugar de tres para evitar posibles confusiones con tipos de texto con nombres similares, tales como *declaration* ("Dec");

(iv) El resto del nombre del archivo aparece tras un espacio en blanco y contiene un breve título que describe el contenido del documento. En este caso, "combating terrorism, "persons traffic supression" y "anti money laundering", respectivamente. En aquellos casos en los que varios archivos contengan un mismo subtítulo, podrán añadirse a éste un índice numérico para diferenciarlos, como por ejemplo "combating terrorism1" y "combating terrorism2". Es importante que en estos casos el primer archivo de la serie numerada contenga el número 1 (y no un espacio en blanco), ya que facilitará su localización y evitará que pueda confundirse con los otros archivos de la misma serie.

Es necesario resaltar que el título de cada archivo no habrá de exceder en ningún caso un máximo de 40 caracteres (espacios incluidos), puesto que el extractor, al que posteriormente se cargará la información, no admite un número mayor de caracteres por título. Por tanto, otra información que pueda resultar relevante en relación con los textos del corpus, como por ejemplo datos relativos a su recopilación o edición, irá incluida en una base de datos y no el nombre del archivo.

3.2 Extracción terminológica

El extractor de FunGramKB permite que la tarea de identificación de términos sea un proceso considerablemente más rápido y efectivo que la inspección manual de las líneas de concordancia de las que se componen los textos. Sin embargo, la decisión sobre qué palabra ha de considerarse un término y cuál ha de descartarse recaerá siempre sobre el terminólogo y dependerá de su criterio. En este sentido, el proceso de extracción en FunGramKB ha de considerarse semiautomático. Para encontrar un grupo de términos representativos de un ámbito de conocimiento es necesario seguir tres pasos fundamentalmente: (i) carga de los textos en el extractor, (ii) extracción de los términos y (iii) selección terminológica. A continuación explicaremos e ilustraremos cada uno de ellos.

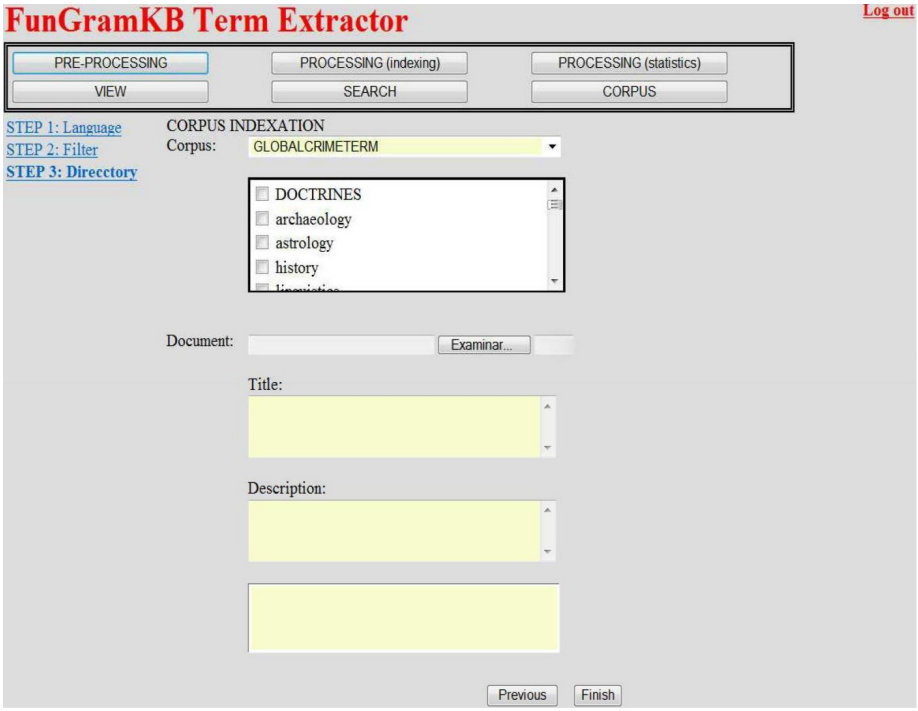
3.2.1 Carga del corpus

La carga del corpus es un proceso que consiste en la introducción progresiva y manual del corpus en el extractor de FunGramKB. Este proceso es, junto con la compilación del corpus y la definición de términos, el paso que requiere mayor precisión y supervisión desde el punto de vista procedimental para garantizar un correcto procesamiento terminológico. Durante la carga, el extractor almacena de forma permanente todo los textos del corpus como paso previo a su procesamiento. Resulta fundamental que, antes de proceder a la carga del corpus, las etapas de recopilación de textos y de edición de archivos hayan concluido, es decir, que el corpus se haya "cerrado" y no se le añada texto alguno a partir de este momento. De esta forma, al empezar la carga, el terminólogo habrá de conocer de forma detallada el número total de textos que se van a cargar, así como el tipo de documentos que contiene el corpus. Según hemos mencionado, el extractor está basado en cálculos estadísticos, por los que la base cuantitativa ha de ser definida e invariable antes de proceder a la

extracción. De ahí que sea imprescindible no proceder a la carga sin haber concluido las etapas previas.

La carga del corpus se hará mediante la herramienta "corpus (indexing)" del extractor y la ejecutará el terminólogo principal, de forma que solamente éste tendrá acceso a la herramienta de indizado. El proceso de carga se divide en dos fases: la "precarga" y la "carga". En la primera de estas fases, el terminólogo, ya situado en la opción "corpus (indexing)", debe introducir cuatro parámetros en el extractor: (i) el texto concreto que va a cargar, (ii) el título del texto (más propiamente, del archivo *txt* que contiene el texto), (iii) una descripción del contenido principal del texto, y d) una etiqueta que identifique el subdominio al que pertenece el texto (por ejemplo, la arqueología, la astrología, o, en el caso del GCTC, el derecho), (Figura 3).

Figura 3. Pantalla de precarga del extractor de FunGramKB



Tanto el título como la descripción no habrán de ser demasiados extensos, y no podrán superar en ningún caso los 40 y 100 caracteres, respectivamente. Una vez seleccionados estos tres parámetros, el terminólogo habrá precargado el texto en el extractor y podrá acceder a la fase de carga, en la que podrá comprobar si el archivo que ha precargado, así como el título y la descripción correspondientes, son los que han de validarse de forma definitiva o, si por el contrario, es necesario hacer alguna rectificación. Para facilitar dicha comprobación, la ventana de carga muestra un resumen en el que incluye el título del texto precargado y su descripción correspondiente. En caso de que sea necesaria alguna corrección, el terminólogo podrá regresar a la fase de precarga y corregir los errores detectados. En el caso de que la precarga haya sido correcta, el terminólogo podrá validar la precarga y el texto pasará a cargarse finalmente en el extractor. Los textos que hayan sido cargados aparecerán reflejados de forma instantánea en la pestaña "corpus" (Figura 4).

Figura 4. Pantalla de carga del extractor de FunGramKB

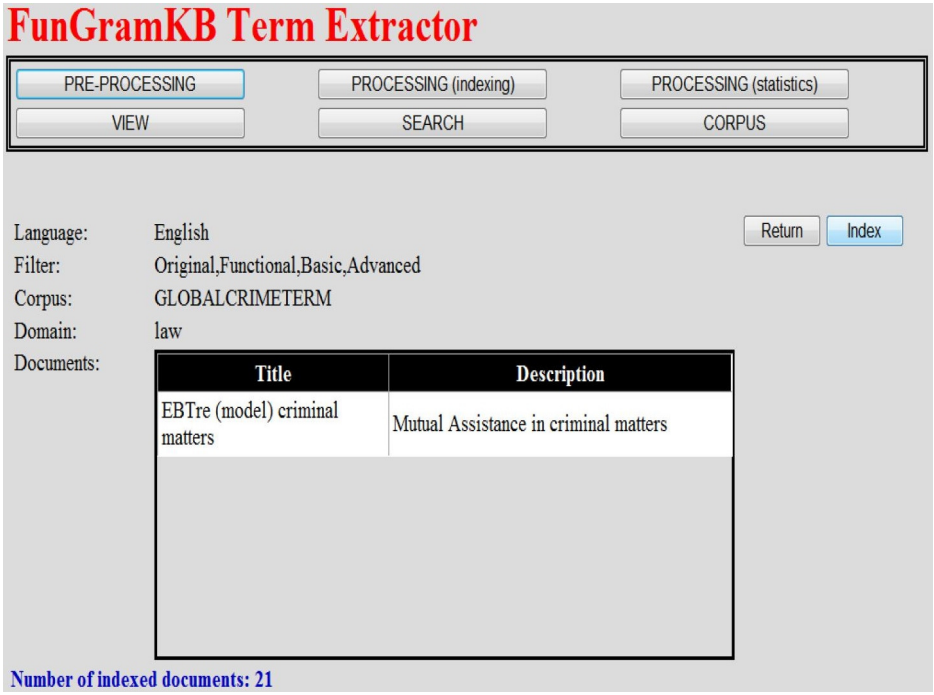
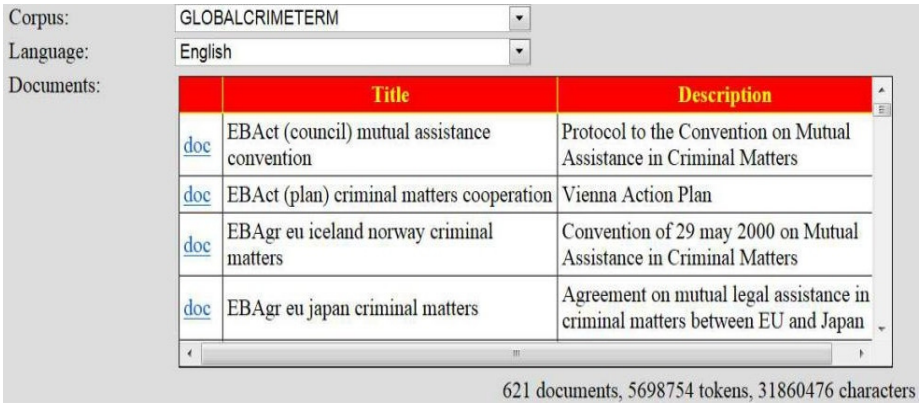


Figura 5. Pantalla de visualización del corpus

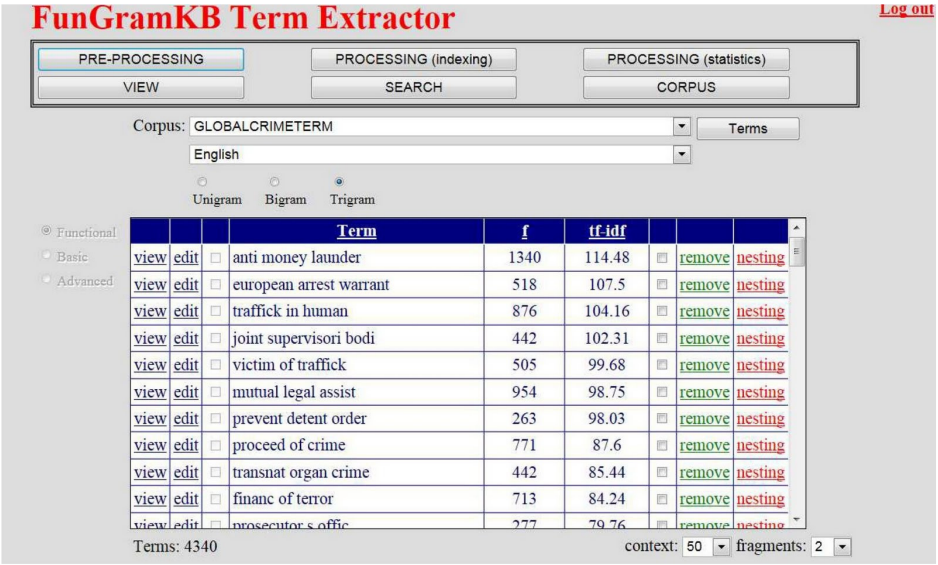


Una de las mayores ventajas que ofrece el extractor es la carga de varios documentos al mismo tiempo, hasta un máximo aproximado de diez archivos simultáneamente, según sea mayor o menor el tamaño global de los archivos. Dado el potencial que esto supone, es crucial que, antes de comenzar, se tenga en cuenta un aspecto muy importante: la carga de un texto no es reversible, es decir, una vez en el extractor, un texto no podrá eliminarse, así como tampoco la información relativa a su título y descripción. Esta situación hace necesario, por tanto, que el terminólogo siga de forma meticulosa el proceso de comprobación de los datos en la etapa de precarga de los textos.

3.2.2 Extracción de términos candidatos

Antes de pasar a la fase de extracción terminológica, es necesario definir algunas nociones centrales del estudio terminológico que serán relevantes en esta fase del proceso. La primera de ellas es la noción de "n-grama", que es una unidad de descripción básica dentro del ámbito del procesamiento computacional del lenguaje. Un n-grama se refiere a una secuencia de caracteres alfabéticos contiguos separada de otras secuencias por un espacio en blanco o un signo de puntuación. Un n-grama, por tanto, coincide con lo que de forma común denominamos "palabra" en su dimensión ortográfica. La letra "n" en "n-grama" es un símbolo matemático que se utiliza para englobar a toda la serie de números naturales, esto es, los números de cero en adelante (1, 2, 3...). Existen tres tipos de n-gramas que se consideran básicos en FunGramKB: los "unigramas", que son palabras individuales (por ejemplo, "crimen"); los "bigramas" o secuencias de dos palabras (por ejemplo, "crimen organizado") y, finalmente, los "trigramas", que son formaciones unificadas de tres palabras (por ejemplo, "crimen organizado internacional"). El extractor de FunGramKB clasifica los términos del corpus en unigramas, bigramas y trigramas, exclusivamente, y son con estos tres tipos de unidades con las que trabajaremos durante el proceso de extracción. La Figura 6 muestra el apartado dentro del extractor de FunGramKB en el que se pueden ver los trigramas, que aparecen listados en la columna denominada "term".

Figura 6. Tabla de trigramas candidatos

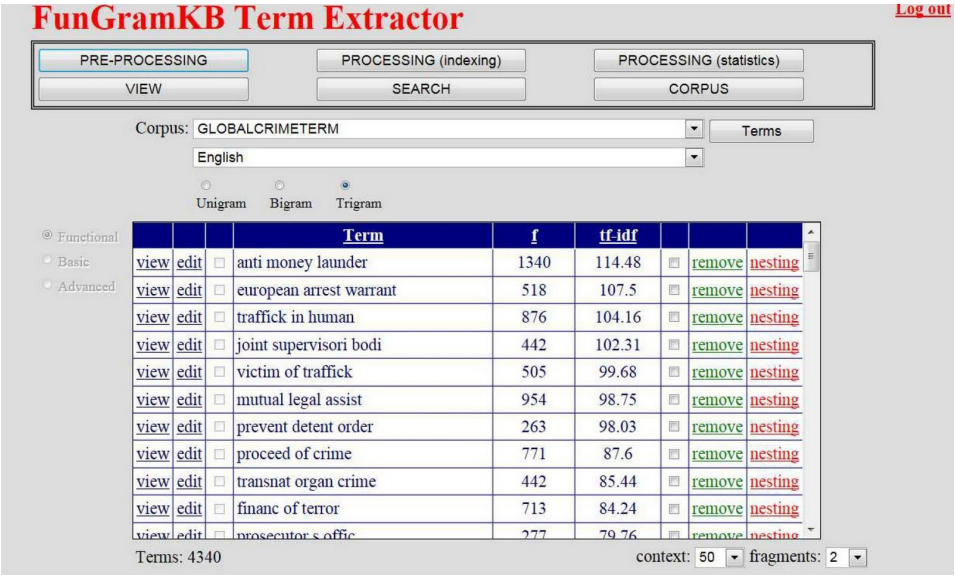


Los n-gramas pueden clasificarse por el número de componentes de los que están formados, pero también por su relevancia terminológica. Según este criterio, pueden distinguirse tres tipos de unidades: "término candidato", "término ganador" y "término falso". Se denomina "término candidato" a un n-grama que, tras un proceso de extracción terminológica, se posiciona como unidad terminológica potencialmente relevante en un dominio léxico concreto. Se denomina "término ganador" a un n-grama candidato que, propuesto por el extractor y ratificado por el terminólogo, se considera un término propio de un ámbito de conocimiento concreto. Finalmente, en contraposición a los términos ganadores, un "término falso" se refiere a un n-grama candidato que, tras ser analizado por el terminólogo, pasa a considerarse una unidad léxica del lenguaje general y, por tanto, no especializado. Otros términos falsos contienen unidades o combinaciones que no están conectadas o simplemente no tienen sentido. El terminólogo encargado del análisis de términos candidatos utilizará

el extractor de FunGramKB con el objetivo de encontrar el mayor número de términos ganadores representativos del dominio de estudio, puesto que sólo estos (y no los términos falsos) serán finalmente definidos y conceptualizados para que cuelguen de la subontología y de los respectivos lexicones, según proceda. En el análisis del GCTC encontramos diversos ejemplos de estos tipos de unidades. Veamos algunos de ellos.

Sólo para el inglés del GCTC se han cargado en el extractor aproximadamente 600 textos. Una vez concluido el proceso de carga, se ha procedido al cálculo estadístico para extraer los términos candidatos. Tras el proceso estadístico, en el que el extractor calcula el índice *tf-idf* (véase apartado 2 arriba) de todas las unidades del corpus, se ha obtenido la lista de trigramas que aparece en la Figura 7.

Figura 7. Tabla de trigramas candidatos



Igualmente, la extracción estadística ha arrojado resultados en los apartados de bigramas y unigramas. Estas unidades candidatas aparecen en las Figuras 8 y 9, respectivamente.

Figura 8. Tabla de bigramas candidatos

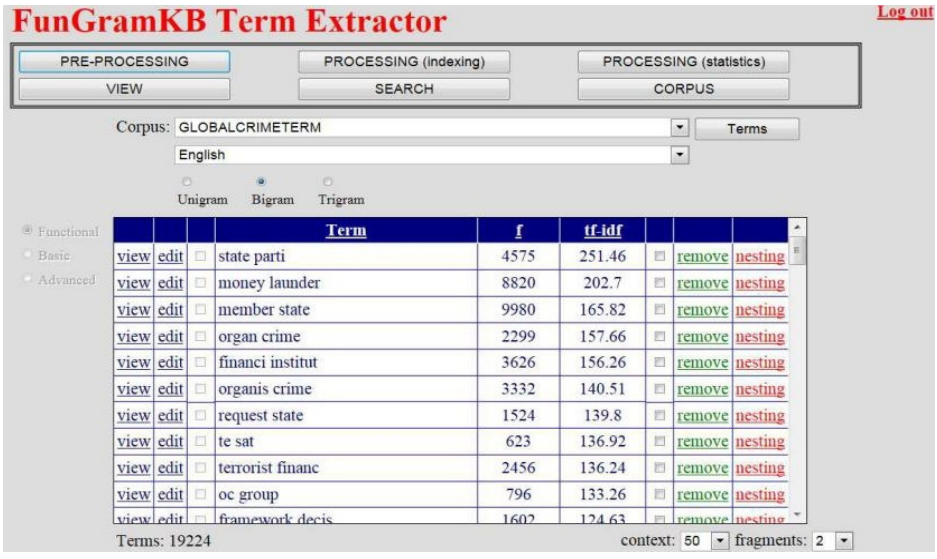
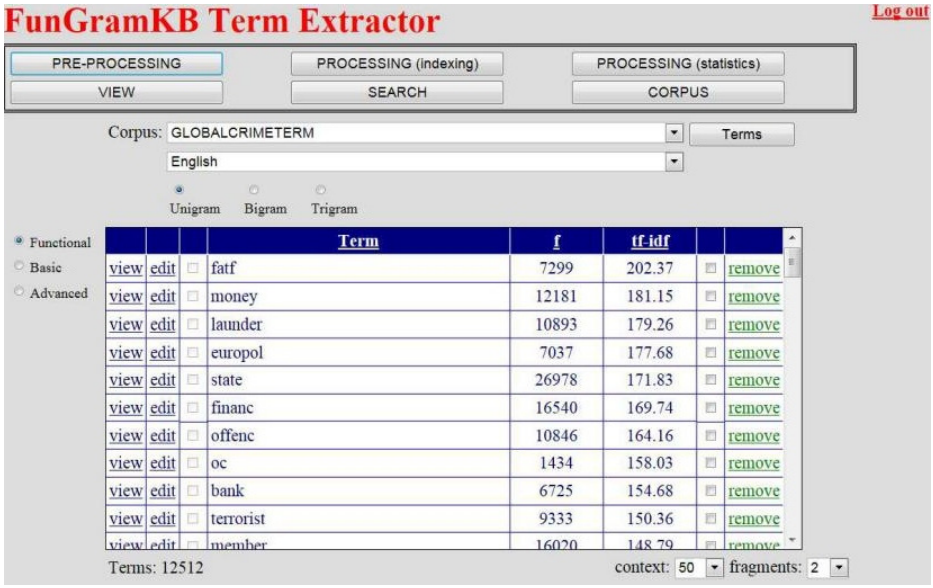


Figura 9. Tabla de unigramas candidatos



De forma previa al cálculo estadístico, el extractor aplica sobre los términos que va procesando un filtro denominado *stemming*, que reduce las unidades que procesa a sus raíces léxicas. De hecho, los n-gramas que visualizamos en las listas de candidatos una vez concluida la extracción son en realidad ngramas truncados o raíces léxicas. Por ejemplo, el unigrama candidato "prevent" es una raíz que agrupa unidades como "prevention", "prevents" o "preventing". Otras raíces son más opacas en una primera inspección en lo que se refiere a su similitud con las variantes morfémicas. Por ejemplo, la variedad morfológica del trigrama "use of hi" puede albergar las siguientes variables que están claramente no relacionadas desde el punto de vista semántico: "use of heroin", "use of human", "use of hearsay", "use of high-level", "use of handling" y "use of household". Con la aplicación del *stemming*, el procesador consigue un mayor agrupamiento terminológico y un ajuste estadístico más preciso de los términos que contiene el corpus. No obstante, el extractor de FunGramKB dispone de una opción para visualizar las variantes morfológicas de los elementos truncados por el *stemmer* y facilitar la comprensión de las raíces, especialmente las menos transparentes.

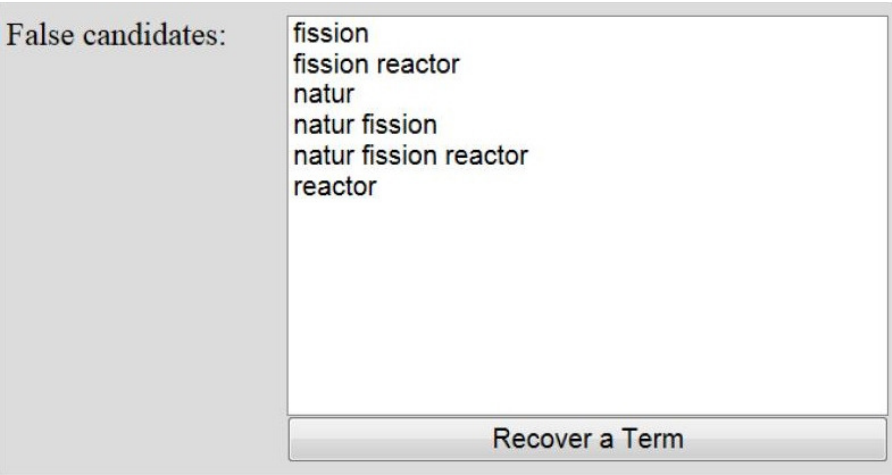
3.2.3 Selección de términos ganadores

Para obtener una lista de términos ganadores es necesario realizar un "filtrado terminológico" de los términos candidatos. Este filtrado consistirá en identificar en la lista inicial de términos candidatos qué términos son realmente terminología y cuáles habrán de ser eliminados por no ser pertinentes. Esta tarea de filtrado la realizará de forma manual el terminólogo y debe considerarse como un paso crucial en la creación subontológica, puesto que de ella dependerá que el modelado conceptual sea representativo, cualitativa y cuantitativamente, del dominio que trata de representar.

Existen dos formas de eliminar un término de la lista de candidatos: el "borrado simple" (*removal*) y el "borrado anidado" (*nesting*). El borrado simple se utiliza para eliminar un único n-grama candidato, cualquiera que sea su complejidad. Por ejemplo, si en la lista de unigramas aparece un candidato "drug" y deseamos que sea descartado por no tratarse de un concepto especializado, emplearemos el borrado simple para eliminarlo de la lista correspondiente. De la misma forma, el borrado simple permite en el caso, por ejemplo, del bigrama candidato "cooperation

agreement", eliminar esta secuencia de dos términos (y sólo esta secuencia), y sólo en el orden en que aparecen dentro del bigrama. El borrado anidado aparecerá como opción exclusivamente en la ventana de bigramas y trigramas candidatos, y será de gran utilidad especialmente en el filtrado de estos últimos. Al seleccionar un término y optar por el borrado anidado, el extractor descartará el trigrama completo y todos sus componentes de forma individual. Por ejemplo, si seleccionamos el candidato "criminal money laundering" y seleccionamos *nesting*, el extractor borrará la secuencia de estas tres palabras pero, a diferencia del borrado simple, también eliminará individualmente "criminal", "money" y "laundering" de toda la lista de n-gramas candidatos. El anidamiento, además de borrar la combinación "criminal money laundering", así como estas tres palabras individualmente, también elimina los emparejamientos entre las unidades que lo componen, esto es, "criminal money" y "money laundering". Es necesario subrayar que, como resultado del borrado anidado, los términos eliminados tampoco aparecerán ya reflejados en la lista de candidatos bigramas o unigramas. En este caso, "criminal", "money" y "laundering" desaparecerían de todas las listas de candidatos. Veamos otro ejemplo. Si en la lista de candidatos aparece el trigrama "natur fission reactor" y el terminólogo considera que no es un trigrama relevante, ni tampoco los unigramas que lo forman, puede utilizar el borrado anidado, en cuyo caso el resultado de la eliminación será el siguiente (Figura 10).

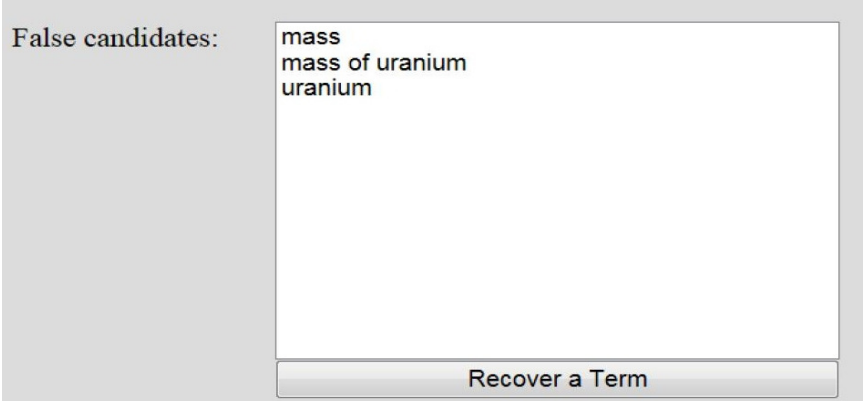
Figura 10. Papelera de reciclaje terminológica. Eliminación de un trigrama



El trigrama "natur fission reactor" aparece en la parte central de la que denominaremos "papelera terminológica". Junto a este trigrama aparecen todas las posibles combinaciones, igualmente eliminadas, de los unigramas y bigramas que lo forman. Como resultado de la eliminación anidada, ninguna de estas unidades aparecerá ya en el resto de listas de candidatos.

Hay que señalar que el borrado anidado sólo filtrará unidades léxicas, es decir, no eliminará palabras funcionales, tales como las preposiciones o el marcador verbal *to*, ya que son parte integrante de otras formaciones trigramáticas y bigramáticas. El filtrado de las palabras funcionales tendrá lugar en la fase de filtrado de unigramas y lo realizará el extractor de FunGramKB de forma automática. Para ilustrar el comportamiento del borrado anidado con respecto a las palabras funcionales, consideremos el trigrama "mass of uranium" (Figura 11), cuyo esquema lingüístico es "léxico + funcional + léxico".

Figura 11. Papelera de reciclaje terminológico. Eliminación de unidades funcionales



Como se observa en la Figura 11, el anidamiento del trigrama "mass of uranium" no ha incluido como término falso la preposición "of", de ahí que ésta no aparezca como entrada individual en la papelera de reciclaje. Esta preposición, por tanto, seguirá apareciendo tanto en la lista de trigramas, como en la de bigramas y unigramas.

Al utilizar el borrado anidado es fundamental que los terminólogos conozcan exactamente la repercusión de este borrado y su alcance sobre cada una de las listas de candidatos del corpus. Supongamos, finalmente, que nos hallamos ante un trigrama que no es relevante, ni tampoco sus componentes, y, en consecuencia, aplicamos un borrado anidado. Como resultado, según se ha explicado anteriormente, los términos formantes desaparecen de la lista de unigramas. Sin embargo, si alguno de estos unigramas es parte integrante de un bigrama, dicho término seguirá en este bigrama y aparecerá en la lista de candidatos bigramas. Por ejemplo, si borramos con anidamiento el trigrama "earth s atmosphr", en la lista de bigramas seguirá apareciendo la combinación "archean atmospher", a pesar de que "atmospher" ha sido eliminado de los trigramas. "Atmospher" no aparecerá ya, no obstante, en la lista de unigramas. Cabe recordar aquí que, tal y como se mencionó en el apartado anterior, el terminólogo trabajará con ngramas formalmente expresados como raíces léxicas, por lo que durante el filtrado éste deberá inspeccionar de forma concreta cada una de las posibles realizaciones morfológicas de los candidatos para minimizar la posibilidad de eliminar realizaciones concretas de la raíz que puedan ser de importancia terminológica.

La Tabla 1 contiene un resumen de todos los casos que pueden ocurrir durante el filtrado terminológico, así como las decisiones que se han de adoptar en cada caso (el símbolo '>' indica "aplíquese", mientras que un asterisco '*', indica "término eliminado", y 'N/A' indica que no es un caso posible en FunGramKB):

(i) Trigrama XYZ (léxico + léxico + léxico):
> *remove* = *XYZ
> *nesting* = *XYZ | *X | *Y | *Z | *XY | *YZ
Ejemplo:
Trigrama "international global crime"
> *remove* = *international global crime
>*nesting* = *international global crime | *international | *global | *crime | *international global | *global crime

(ii) Trigrama XYZ (léxico + funcional + léxico)
> *remove* = *XYZ
> *nesting* = *XYZ | *X | *Z
Ejemplo:

Trigrama “crimin or terrorist” (p. ej., “criminal or terrorist”)
> *remove* = *crimin or terrorist
> *nesting* = *crimin or terrorist | *crimin | *terrorista

(iii) Trigrama XYZ (léxico + léxico + funcional)
> *remove* = *XYZ
> *nesting* = *XYZ | *X | *Z
Ejemplo:
Trigrama “access to such”
> *remove* = *access to such”
> *nesting* = *access to such | *access | *such

(iv) Trigrama XYZ (funcional + léxico + léxico)
N/A

(v) Bigrama XY (léxico + léxico)
> *remove* = *XY
> *nesting* = *XY | *X | *Z
Ejemplo:
Bigrama “avoid transact” (p. ej.,“avoid transaction”)
> *remove* = *avoid transact
> *nesting* = *avoid transact | *avoid | *transact

(vi) Bigrama XY (léxico + funcional)
N/A

(vii) Bigrama XY (funcional + funcional)
N/A

(viii) Unigrama X (léxico)
> *remove* = *X
> repercusión: al borrar el unigrama X, X ***no*** desaparece de las formaciones bigramáticas o trigramáticas que contengan X.

Tabla 1. Metodología de filtrado terminológico

El procedimiento óptimo para el filtrado de candidatos es el denominado procesamiento descendente, es decir, el terminólogo habrá de decidir, en primer lugar, qué trigramas son ganadores y cuáles no, y sólo una vez haya terminado este primer filtrado, pasará a filtrar los bigramas y los unigramas, por este orden. Esta metodología permite optimizar el tiempo global de filtrado del corpus completo, ya que el extractor, tal y como se acaba de señalar respecto al borrado anidado, permite el borrado múltiple de forma simultánea en las tres listas de candidatos, lo que hace que el filtrado resulte más rápido. Una vez llegado a la etapa de filtrado de unigramas, el extractor ofrece la posibilidad de aplicar tres filtros automáticos a la lista de candidatos. En este caso el terminólogo deberá empezar el filtrado, en primer lugar, con el filtro avanzado activado (“advanced”). Este filtro criba miles de palabras de uso muy frecuente propias de los diccionarios no especializados. Una vez concluido este filtrado, el terminólogo habrá de proceder a seleccionar el filtro básico (“basic”), en el que la lista de candidatos queda automáticamente desprovista de unos cientos de términos no especializados de uso muy frecuente. Finalmente, el filtro funcional (“functional”) quitará de la lista de candidatos unidades funcionales como los pronombres, los auxiliares verbales o las preposiciones, ya que, por sí solas, estas unidades nunca constituirán un término. Cabe recordar, finalmente, que la eliminación de unigramas, en cualquiera de estos tres estadios de filtrado, no supone su eliminación de bigramas o trigramas, es decir, no existe anidamiento posible.

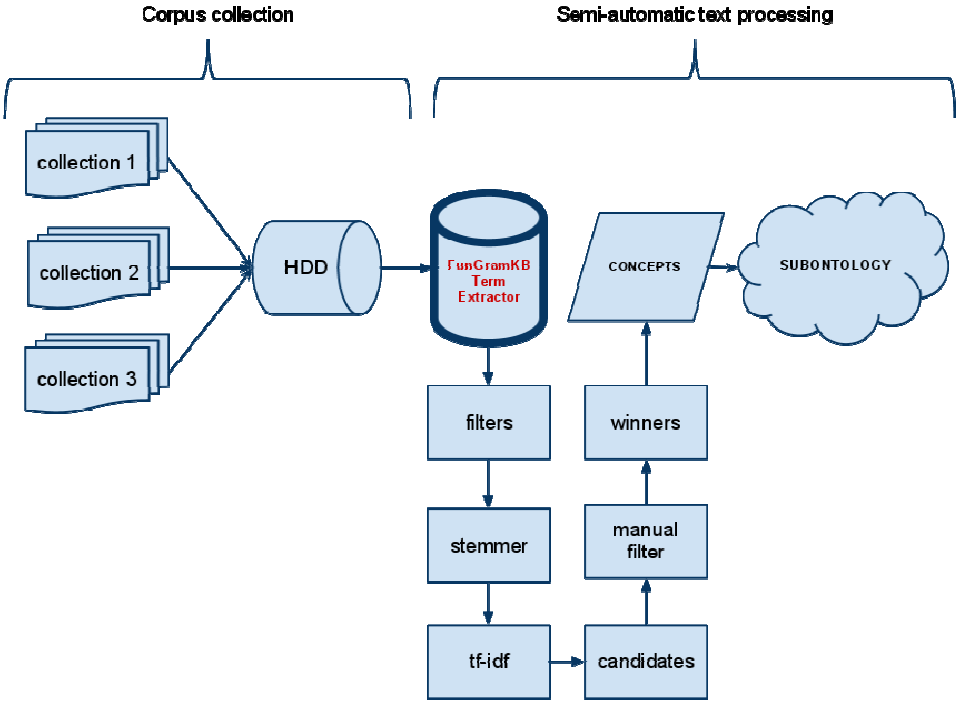
Los términos falsos que han resultado eliminados durante el filtrado quedarán almacenados de forma provisional dentro del extractor en la sección “corpus”. Allí

permanecerán listados durante el proceso de filtrado hasta el momento en que se valide la lista definitiva de términos ganadores, momento en el que el extractor purgará automáticamente la papelera terminológica. Hasta que tenga lugar la validación, los términos eliminados estarán accesibles en todo momento, de forma que, en caso que el terminólogo así lo decida, podrán recuperarse para reubicarse automáticamente en la correspondiente lista de n-gramas candidatos de la que fueron eliminados. En el caso de que un término haya sido eliminado mediante anidamiento, la recuperación se hará de forma individualizada de cada una de las combinaciones formantes que fueron eliminadas como resultado, y no de todas ellas de forma conjunta. El proceso de recuperación puede tener lugar tantas veces como sea necesario y con el número de términos distintos que se requiera. El extractor, por tanto, es muy flexible a la hora de administrar los términos candidatos y su borrado.

3.3 Definición terminológica

En los apartados anteriores hemos analizado cómo extraer una lista de términos especializados de un dominio a partir de un corpus de ejemplos. Estas dos fases, así como las subtareas que se realizan en cada una de ellas, podrían resumirse según muestra la Figura 12.

Figura 12. Proceso de compilación y filtrado en FunGramKB



El presente apartado se centrará en la etapa final de la adquisición terminológica. Abordaremos el tratamiento de los términos ganadores y el uso de diccionarios para la definición de los mismos.

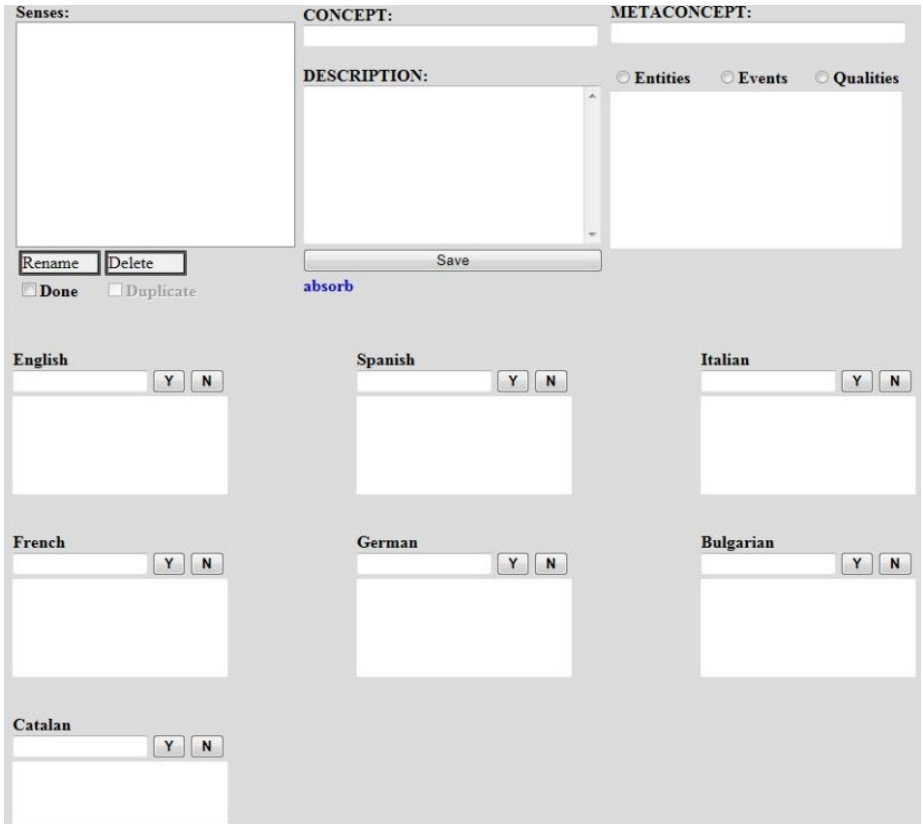
3.1.1 Edición de términos

Una vez obtenidos los términos que consideramos ganadores, el siguiente paso consistirá en definirlos de manera adecuada. Para ello, habrá que establecer una serie de consideraciones previas, como son la organización de la estructura conceptual

nuclear de la disciplina seleccionada a partir de las evidencias obtenidas en los marcos epistemológicos de ontologías precedentes (si existen) y también a partir del asesoramiento de expertos y fuentes consultadas por el equipo investigador. Esas estructuras fundamentadas en una aproximación deductiva al dominio temático seleccionado se irán verificando o sustituyendo mediante la fase metodológica inductiva del estudio. Dicho de otra forma, el trabajo de campo ratificará o no la coherencia o solvencia de esa estructura propuesta.

La identificación de las palabras definitorias del dominio temático elegido a partir de búsquedas en los recursos y fuentes disponibles en formato electrónico, tal y como ha quedado expuesto anteriormente, permitiría el volcado de la información en FunGramKB a través de su editor on-line (véase la Figura 13), que conectará la subontología que se cree con la Ontología Nuclear (plano conceptual), de una parte, y con los lexicones de las lenguas seleccionadas (plano lingüístico), de otra.

Figura 13. Herramienta de edición de términos ganadores



3.3.1.1 Selección de diccionarios y fuentes lexicográficas

La selección de los diccionarios es una tarea crucial para la extracción de una definición apropiada. Ahora bien, debe tenerse en cuenta que los diccionarios no forman parte del propio corpus, sino que se utilizan como herramienta de apoyo al terminólogo para construir las definiciones de los términos que se identifiquen en el corpus. El propósito de los diccionarios es asistir al experto en la elaboración de la definición de los términos, pero no en la identificación de los mismos. En el marco de FunGramKB es fundamental tener en cuenta que las definiciones de los términos, independientemente de la lengua utilizada, se escribirán todas ellas en inglés, aunque se traten de términos específicos de una sola lengua. La razón es doble: (i) el inglés es la lengua de interfaz de FunGramKB, y (ii) el vocabulario definitorio básico se obtendrá de estas definiciones, por lo cual todas las definiciones de los términos en cualquiera

de las lenguas elegidas deben estar escritas en el mismo idioma. Por ejemplo, aunque se trate de un término especializado que sólo se utilice en italiano, y no tenga traducción en otros idiomas, su definición se compilará en inglés.

Dado que lo usual es operar con varias lenguas al mismo tiempo en lugar de una única lengua en este tipo de actividades lexicográficas, una manera recomendable de proceder es seleccionar los que se consideren como tres mejores diccionarios de la especialidad con los perfiles siguientes:

(i) Tres diccionarios monolingües para cada lengua (preferiblemente en formato electrónico, si existen)
(ii) Tres diccionarios bilingües (preferiblemente en formato electrónico, si existen)

3.3.1.2 Elaboración de las definiciones

Una vez completada la primera tarea de la obtención del repositorio terminológico especializado (tarea semiautomática), el siguiente cometido es organizar jerárquicamente los términos de ese repositorio (tarea asistida) a través de la relación taxonómica de la subsunción (IS-A). En esta segunda etapa, nos basaremos principalmente en los diccionarios, ya que en los textos definitorios se explicita el término superordinado del *definiendum*.

Tomemos el término +CRIME_00 como ejemplo de la construcción de una subontología sobre crimen organizado. Posiblemente, este superordinado (o hiperónimo) puede (i) pertenecer al subdominio del crimen organizado, (ii) pertenecer al dominio legal general, o (iii) no pertenecer a la ontología satélite legal, sino que se trate de un concepto básico/terminal de la Ontología Nuclear (i.e. no terminológico). Como puede observarse, su presencia en la Ontología Nuclear se clasifica de hecho como caso (iii). Véanse las Figuras 14 y 15.

Figura 14: Información conceptual de +CRIME_00 en la Ontología Nuclear de FunGramKB

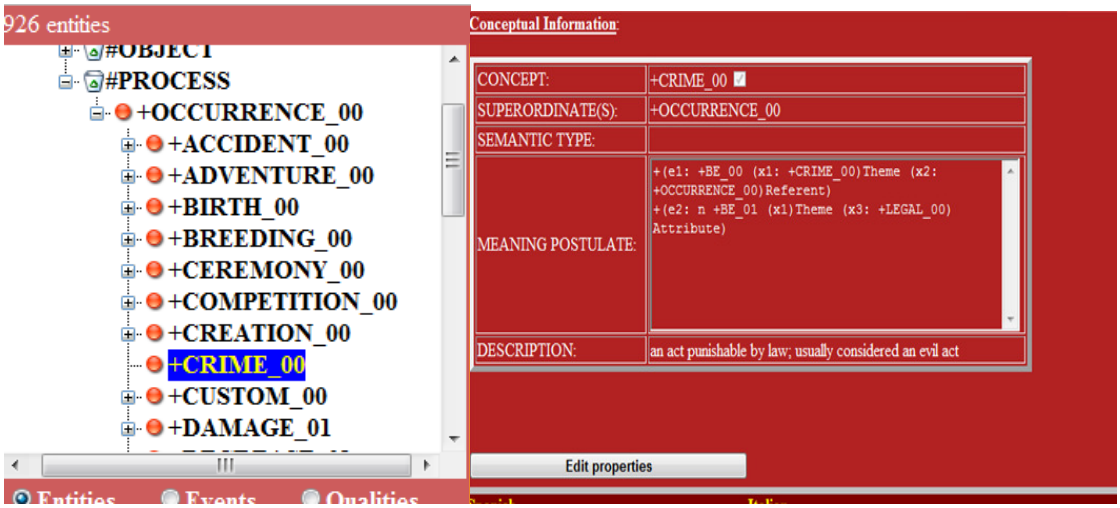
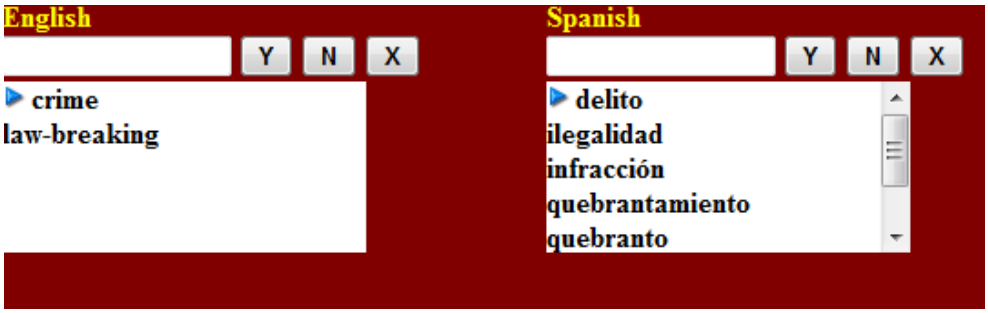


Figura 15: Representación léxica del concepto +CRIME en la Ontología Nuclear de FunGramKB



Otros términos más especializados a partir de +CRIME_00, como podría ser (a) +FELONY_00 (posiblemente, caso ii) o (b) $WATERBOARD (posiblemente, caso i)[7], una vez confirmada su presencia en el corpus como términos ganadores, habría que determinar finalmente si se trata de conceptos básicos o terminales en una subontología especializada en crimen organizado, dado que en ningún caso formarían parte de la Ontología Nuclear. Esto lo va a determinar el estudio detallado de los elementos que integrarían la definición de acuerdo con las fuentes consultadas.

Esta diversidad no es en realidad un problema, sino simplemente es importante conocer qué lugar corresponde a cada concepto, término, etc. El caso (i) es muy simple, ya que el término está ampliamente especificado en nuestro repositorio terminológico. El caso (ii) es igualmente sencillo, ya que implicaría introducir en la ontología satélite un nuevo concepto y descubriríamos cuál es a su vez el superordinado hasta llegar a un concepto que se encuentre en la Ontología Nuclear (+FELONY_00 está subordinado a +CRIME_00 u +OFFENCE_00, y estos a su vez se subordinan a +OCCURRENCE_00). Esto es lo más importante, de hecho. Conectar la ontología satélite con la Ontología Nuclear común. El caso (iii) es uno de los más sencillos, ya que esa conexión se explicita desde el principio.

4 Conclusiones

Este artículo ha presentado los aspectos más relevantes de la metodología de extracción terminológica dentro del marco de FunGramKB y algunos principios básicos y esquemáticos de la definición terminológica. La adquisición terminológica conforma la base sobre la que se fundamentan labores de tanta relevancia como el procesamiento textual, la minería de datos o la creación de subontologías especializadas. El artículo ha descrito *FunGramKB Term Extractor* como una herramienta eficiente para la recuperación de terminología cuya función consiste en asistir al terminólogo en todo el proceso de recuperación de términos especializados. El extractor obtiene listas de términos a partir de un corpus de textos de un dominio profesional o técnico. El extractor permite tanto la aplicación de filtros lingüísticos para la automatización de tareas comunes, como la supresión de términos funcionales, sin olvidar la aplicación de cálculos estadísticos para la obtención de una lista de términos especializados que son relevantes en un área de estudio concreta. La extracción de términos basada en corpus abre las puertas a la creación de subontologías más

[7] El proponer en este caso la hipótesis de que +FELONY_00 pueda ser un concepto básico (asignación del signo +) y $WATERBOARD uno terminal (asignación del signo $) en una ontología satélite sobre crimen organizado es algo que en un primer análisis parece probable pero que habrá que verificar en una fase ulterior del proyecto.

completas y mejor definidas, y supone un salto cualitativo en el conocimiento especializado y en el modelado conceptual del mundo.

Agradecimientos

Esta contribución forma parte del proyecto de investigación denominado *Elaboración de una subontología terminológica en un contexto multilingüe (español, inglés e italiano) a partir de la base de conocimiento FunGramKB en el ámbito de la cooperación internacional en materia penal: terrorismo y crimen organizado*, financiado por el Ministerio de Ciencia e Innovación. Código: FFI2010-15983.

Referencias

Biber, Douglas, Conrad, Susan y Reppen, Randi (1998): *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Cabré, María T. (1999): *Terminology: Theory, Methods, and Applications*. Amsterdam: John Benjamins.

Felices-Lago, Ángel y Marín-Rubiales, Amalia (En prensa): "La semántica profunda como fundamento para desarrollar una subontología jurídica en el contexto de las ontologías de ámbito legal". En: *Actas del XXIX Congreso Internacional de AESLA*.

Kennedy, Graeme D. (1998): *An Introduction to Corpus Linguistics*. Harlow: Addison Wesley Longman.

Koester, Almut (2010): "Building small specialised corpora". En: Anne O' Keeffe y Michael J. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. Londres: Routledge, 66-79.

Mairal-Usón, Ricardo y Periñán-Pascual, Carlos (2009): "The anatomy of the lexicon component within the framework of a conceptual knowledge base". *Revista Española de Lingüística Aplicada* 22, 217-244.

Mairal-Usón, Ricardo y Periñán-Pascual, Carlos (2010): "Role and Reference Grammar and Ontological Engineering". En José L. Cifuentes, Gómez, A., Lillo, A., Mateo, J. y F. Yus (eds.) *Los Caminos de la Lengua: Estudios en Homenaje a Enrique Alcaraz Varó*. Alicante: Universidad de Alicante, 649-665.

Mairal-Usón, Ricardo, Periñán-Pascual, Carlos y Samaniego, Eva (2011): "Using ontologies for terminological knowledge representation: a preliminary discussion". En: *Technological Innovation in the Teaching and Processing of LSPs: Proceedings of TISLID'10*. Madrid: UNED, 267-280.

Meyer, Charles F. (2002): *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Periñán-Pascual, Carlos y Arcas-Túnez, Francisco (2004): "Meaning postulates in a lexico-conceptual knowledge base". En: *15th International Workshop on Databases and Expert Systems Applications,* Los Alamitos (California): IEEE, 38-42.

Periñán-Pascual, Carlos y Arcas-Túnez, Francisco (2005): "Microconceptual-Knowledge Spreading in FunGramKB". En: *Proceedings on the 9th IASTED International Conference on Artificial Intelligence and Soft Computing*. Anaheim-Calgary-Zurich: ACTA Press, 239-244.

Periñán-Pascual, Carlos y Mairal Usón, Ricardo (2009): "Bringing Role and Reference Grammar to natural language understanding". *Procesamiento del Lenguaje Natural* 43, 265-273.

Periñán-Pascual, Carlos y Mairal Usón, Ricardo (2010): "La gramática de COREL: un lenguaje de representación conceptual". *Onomázein* 21, 11-45.

Reppen, Randi (2010): “Building a corpus: what are the key considerations?” En: Anne O’ Keeffe y Michael J. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. Londres: Routledge, 31-37.

Temmerman, Rita (2000): *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam: John Benjamins.

Wright, Sue E. y Budin, Gerhard (1997): *Handbook of Terminology Management: Basic Aspects of Terminology Management.* Vol. 1. Amsterdam: John Benjamins.