Revista Signos
ISSN: 0035-0451
revista.signos@ucv.cl
Pontificia Universidad Católica de Valparaíso
Chile

Periñán-Pascual, Carlos; Arcas-Túnez, Francisco
La ingeniería del conocimiento en el dominio legal: La construcción de una Ontología Satélite en FunGramKB
Revista Signos, vol. 47, núm. 84, marzo-, 2014, pp. 113-139
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Revista Signos. Estudios de Lingüística ISSN 0718-0934
© 2014 PUCV, Chile DOI: 10.4067/S0718-09342014000100006 47(84) 113-139

La ingeniería del conocimiento en el dominio legal: La construcción de una Ontología Satélite en FunGramKB

Knowledge engineering in the legal domain: The construction of a FunGramKB Satellite Ontology

Carlos Periñán-Pascual
Universidad Politécnica de Valencia
España
jopepas3@upv.es

Francisco Arcas-Túnez
Universidad Católica San Antonio
España
farcas@pdi.ucam.edu

Recibido: 15-X-2012 / **Aceptado:** 30-VII-2013

Resumen

Una de las tareas más tediosas en la labor diaria de los profesionales del derecho es la búsqueda de información en el ámbito jurídico. Con el fin de implementar aplicaciones avanzadas del procesamiento del lenguaje natural en este dominio, hemos desarrollado un modelo de representación del conocimiento especializado orientado a la semántica profunda dentro del marco de FunGramKB, una base de conocimiento léxico-conceptual multilingüe de propósito general. Más concretamente, el resultado de esta investigación ha dado como fruto una ontología terminológica sobre derecho penal en el dominio del terrorismo y el crimen organizado transnacional para ser utilizada en sistemas inteligentes que permitan la comprensión automática del discurso legal. El objetivo de este artículo es la descripción de la metodología empleada en el desarrollo de dicha ontología, centrándonos en la descripción de la herramienta que asiste al lingüista en el proceso de adquisición y conceptualización de los términos.

Palabras Clave: Ingeniería del conocimiento, ontología, FunGramKB, terminología, derecho.

Abstract

One of the most time-consuming tasks in the daily work of legal professions is the search for information in the field of law. To implement advanced computer-based applications of natural language processing in this regard, we have developed a model of specialized knowledge representation driven by the deep semantics of FunGramKB, a multilingual general-purpose lexico-conceptual knowledge base. In particular, our research results in a terminological ontology on criminal law in the domain of transnational terrorism and organized crime to be implemented in intelligent systems which aim to understand legal discourse automatically. The objective of this paper is to describe the methodology used in the development of that ontology, focusing on the computerised tool to assist linguists in the process of terminological acquisition and conceptualization.

Key Words: Knowledge engineering, ontology, FunGramKB, terminology, law.

INTRODUCCIÓN

Una ontología en el ámbito de la ingeniería del conocimiento está conformada por una jerarquía de conceptos, atributos y relaciones consensuados que permite establecer redes semánticas de relaciones (Gruber, 1993). La descripción de ontologías como esquemas conceptuales dentro del dominio legal surge en 1995 (Valente & Breuker, 1994; Valente, 1995) por la necesidad de formalizar la conexión y el intercambio de información entre los elementos que conforman un sistema jurídico. Posteriormente, se formalizaría como disciplina de investigación en 1997 con el primer *workshop* sobre ontologías legales dentro del congreso bianual sobre inteligencia artificial y leyes (ICAIL-1997). El objetivo fue y sigue siendo proporcionar mecanismos efectivos para el acceso y la gestión del rápido y creciente volumen de información legal que se produce a diario, actualmente en formato electrónico (Breuker, Casanovas, Klein & Francesconi, 2009). En 2005, André Valente catalogó veinticuatro ontologías en el ámbito legal (Valente, 2005). Luego, Breuker et al. (2009) incrementaron ese listado a treinta y tres, entre las que destacamos FOLaw (Functional Ontology of Law) de Valente y Breuker (1994), FBO (Frame-Based Ontology of Law) de van Kralingen (1995), OPLK (Ontology of the Professional Legal Knowledge) de Benjamins, Contreras, Casanovas, Ayuso, Becue, Lemus y Urios (2004), AGO (Portuguese Attorney General's Office Ontology) de Saias y Quaresma (2003), Jur-Wordnet de Gangemi, Sagri y Tiscornia (2003), DALOS de Francesconi y Tiscornia (2008) y OPJK (Ontology of Professional Judicial Knowledge) de Casanovas, Casellas y Vallbé (2009). Todas estas ontologías vienen expresadas en las principales lenguas de la Unión Europea (por ejemplo: alemán, español, francés, inglés, italiano y portugués), aunque el idioma predominante es el inglés y tan solo una es multilingüe (DALOS).

Además, todas ellas están definidas o bien para un propósito general o bien para un determinado dominio, pero solo una está modelada en ambos niveles. En cuanto

a su construcción, dieciocho se crearon de forma manual (por ejemplo: FOLaw, FBO, Jur-Wordnet y OPJK), dos de forma automática (por ejemplo, AGO) y el resto de forma semiautomática (por ejemplo, OPLK). Las aplicaciones de estos proyectos se centraron en la comprensión del dominio del discurso, el razonamiento y la resolución de problemas, la organización y estructuración de la información, la indización y búsqueda semántica, la digitalización y clasificación de contenidos audiovisuales, y la indización multilingüe. Casanovas, Sartor, Biasiotti y Fernández-Barrera (2011) presentaron una tabla comparativa de estas ontologías. En la actualidad, podemos contar con más de sesenta ontologías legales, junto con diversas tesis doctorales y proyectos financiados en curso.

El desarrollo ontológico puede convertirse en una tarea tediosa que consume mucho tiempo y esfuerzo cuando se realiza de manera completamente manual. Por otra parte, la adquisición automática del conocimiento ontológico parece no haber alcanzado suficiente madurez en las investigaciones, a pesar del éxito relativo de proyectos como OntoLearn (Velardi, Navigli, Cuchiarelli & Neri, 2005) o Text2Onto (Cimiano & Völker, 2005). Actualmente, el paradigma más fiable para el aprendizaje de ontologías recurre con frecuencia a métodos semiautomáticos (Cimiano, Mädche, Staab & Völker, 2009).

Cabe destacar la escasez de ontologías semánticas multilingües construidas de forma semiautomática para garantizar un equilibrio entre la cantidad y la calidad del conocimiento. Además, la adecuación de una ontología depende estrechamente de la aplicación específica para la que se desarrolle. A este respecto, el propósito de nuestra ontología terminológica es su aplicación en sistemas que requieran la comprensión del lenguaje natural, cuyo objetivo es la 'interpretación' de un texto de entrada: es decir, convertir el *input* en una representación formal no ambigua que exprese el contenido semántico del texto con el fin de poder realizar posteriores tareas en diversas aplicaciones del procesamiento del lenguaje natural (PLN), tales como los sistemas de resumen automático o las interfaces persona-máquina basadas en el diálogo (Ovchinnikova, 2012). Desde los inicios de nuestro proyecto (Periñán-Pascual & Arcas-Túnez, 2004, 2005), el cual se enmarca en los campos de la lingüística, la ingeniería del conocimiento y la inteligencia artificial, nuestra intención siempre ha sido el desarrollo de un modelo de conceptualización tratable por la máquina que sirviera para simular el razonamiento humano. En este escenario, la base de conocimiento debe estar integrada en una arquitectura cognitiva, por ejemplo, CLARION (Sun, 1997, 1999; Sun, Merrill & Peterson, 2001), la cual determine la infraestructura del sistema inteligente. Uno de los problemas más frecuentes se encuentra en el hecho de que las arquitecturas cognitivas suelen tratar el conocimiento en forma de categorías opacas, en lugar de representar su significado explícitamente (Langley, Laird & Rogers, 2009). Por tanto, nuestra base de conocimiento, y por ende nuestra ontología terminológica sobre derecho penal en el dominio del terrorismo y el crimen organizado, pretende contribuir a la construcción de sistemas más eficientes para la comprensión textual.

Con respecto a la calidad del conocimiento semántico, el significado de las unidades léxicas puede ser descrito por medio de rasgos o primitivos semánticos o a través de asociaciones con otras unidades léxicas (Velardi, Pazienza & Fasolo, 1991). Por tanto, existe una clara dicotomía entre la semántica profunda, la cual está basada en el significado conceptual, y la semántica superficial, la cual está basada en el significado relacional. En sentido estricto, esta última no proporciona una verdadera definición de las unidades léxicas, sino más bien se limita a describir el uso de las palabras por medio de relaciones de significado con otras unidades léxicas. Por ejemplo, Jur-Wordnet (Sagri, Tiscornia & Bertagna, 2004; Gangemi et al., 2005) se adscribe al enfoque relacional, ya que se trata de una extensión para el dominio legal de la versión italiana de EuroWordNet (Vossen, 1998), el cual se basa a su vez en el modelo léxico de WordNet (Fellbaum, 1998).

Por ello, los *synsets*[1] en Jur-WordNet se relacionan unos con otros a través de relaciones semánticas tales como hiperonimia, hiponimia, meronimia, papel temático o instancia-de[2]. En cambio, algunos ejemplos de ontologías que adoptan un enfoque más profundo son FBO (van Kralingen, 1995; Visser, 1995; van Kralingen, Visser, Bench Capon & van den Herik, 1999), el cual se centra en la noción de marco que introdujo Minksy (1975) en inteligencia artificial, y FrameNet para el dominio legal (Venturi, Lenci, Montemagni, Vecchi, Sagri, Tiscornia & Agnoloni, 2009), el cual representa el conocimiento de acuerdo con la semántica de marcos de Fillmore (1982, 1985).

Así, la ontología legal genérica en FBO distingue tres tipos de entidades: normas, actos y descripciones conceptuales. La ontología define un marco para cada una de estas entidades, el cual especifica los atributos más relevantes de la entidad en cuestión. Por ejemplo, los actos (i.e. eventos y procesos) se representan formalmente a través de marcos que contienen atributos como Agente, Causa, Circunstancia, Medio, Modo, o Tiempo entre otros. Por otra parte, en nuestro segundo ejemplo, los elementos presentes en el repositorio general de FrameNet fueron clasificados a través de una serie de tipos semánticos. No obstante, estos dos casos de ontologías orientadas a los marcos presentan un enfoque engañosamente profundo de la representación del conocimiento especializado, ya que la descripción del significado tiene lugar en el dominio conceptual por medio de una lista de papeles (o elementos del marco) que funcionan en definitiva como relaciones semánticas binarias. Por tanto, se trata de un formalismo de representación del conocimiento cuyo poder expresivo está extremadamente restringido, lo cual repercute en la calidad del conocimiento semántico, ya que los formalismos de representación imponen sus debilidades sobre la axiomatización final (Hakimpour & Geppert, 2001).

Por consiguiente, tanto FBO como FrameNet para el dominio legal adoptan un enfoque ubicado en algún punto intermedio de un *continuum* cuyos extremos son el enfoque superficial de Jur-WordNet y un enfoque marcadamente profundo como el

nuestro, el cual se describe en el próximo apartado. A pesar de que el desarrollo a gran escala de recursos dotados de semántica profunda requiere mucho más tiempo y esfuerzo, su poder expresivo no es solo más robusto, sino además la gestión del conocimiento resulta más eficiente, como demostraron Periñán-Pascual y Arcas-Túnez (2007a). En algunos sistemas del PLN, por ejemplo, la indización automática o la extracción de información, la semántica superficial puede resultar suficiente, pero la construcción de una base de conocimiento robusta garantiza su uso en la mayoría de tareas para el PLN, permitiendo así la reutilización de los recursos.

El resto de este artículo se estructura de la siguiente manera: el apartado 1 presenta brevemente la arquitectura de nuestra base de conocimiento, centrándonos principalmente en los modelos ontológicos; el apartado 2 introduce el marco metodológico que nos permitió construir la ontología de conocimiento especializado; y finalmente, el apartado 3 describe la herramienta informática que asiste al lingüista en las tareas de adquisición y conceptualización de los términos.

1. FunGramKB

1.1. Arquitectura de la base de conocimiento

FunGramKB[3] (Periñán-Pascual & Arcas-Túnez, 2007b, 2010; Mairal-Usón & Periñán-Pascual, 2009; Periñán-Pascual & Mairal-Usón, 2010) es una base de conocimiento léxico-gramático-conceptual multipropósito diseñada principalmente para su uso en sistemas del PLN y, más concretamente, para aplicaciones que requieran la comprensión del lenguaje. Por una parte, esta base de conocimiento es 'multipropósito' en el sentido de que es tanto multifuncional como multilingüe. De esta manera, FunGramKB ha sido diseñada con el fin de ser potencialmente reutilizada en diversas tareas del PLN (por ejemplo: recuperación y extracción de información, traducción automática, sistemas basados en el diálogo, etc.) y con diversas lenguas (alemán, búlgaro, catalán, español, francés, inglés e italiano). Por otra parte, FunGramKB comprende tres niveles principales de conocimiento (i.e. léxico, gramatical y conceptual), cada uno de los cuales está constituido por diversos módulos independientes aunque claramente interrelacionados:

Nivel léxico:

(i) El Lexicón almacena principalmente información morfosintáctica sobre las unidades léxicas.

(ii) El Morficón asiste tanto al analizador como al generador en el tratamiento de los casos de morfología flexiva.

Nivel gramatical:

(iii) El Gramaticón, el cual se estructura siguiendo las directrices del Modelo Léxico-Construccional (Ruiz de Mendoza & Mairal-Usón, 2008; Mairal-Usón & Ruiz de Mendoza, 2009), almacena los esquemas construccionales que pueden ser utilizados por el algoritmo de enlace sintáctico-semántico de la Gramática del Papel y la Referencia (Van-Valin & LaPolla, 1997; Van-Valin, 2005).

Nivel conceptual:

(iv) La Ontología se presenta como una jerarquía IS-A de unidades conceptuales, las cuales contienen el conocimiento semántico en forma de postulados de significado. El modelo ontológico, el cual permite la herencia múltiple no monotónica[4], consiste en dos tipos de componentes: un módulo de propósito general (i.e. Ontología Nuclear) y varios módulos de conocimiento especializado (i.e. Ontologías Satélites).

(v) El Cognicón almacena el conocimiento procedimental por medio de guiones, i.e. esquemas conceptuales que describen una serie de eventos prototípicos dentro de un marco temporal, más concretamente adoptando el modelo temporal de la lógica de intervalos de Allen (1983). Los guiones nos permiten describir, por ejemplo, cómo se hace una tortilla o cómo se realiza una compra online.

(vi) El Onomasticón almacena el conocimiento cultural sobre instancias de entidades y eventos, tales como Cervantes o el 11-M. Este módulo almacena su conocimiento por medio de dos tipos diferentes de esquemas (i.e. retratos e historias), ya que las instancias pueden ser descritas sincrónica o diacrónicamente.

En la arquitectura de FunGramKB, cada lengua tiene sus propios módulos léxico y gramatical, mientras que cada módulo conceptual es compartido por todas las lenguas. En otras palabras, los lingüistas deben construir un Lexicón, un Morficón y un Gramaticón para el español, y lo mismo para cada una de las lenguas restantes, pero los ingenieros del conocimiento solo necesitan construir una Ontología, un Cognicón y un Onomasticón para procesar conceptualmente un texto de entrada. Ya que la Ontología se convierte en el pivote de toda la arquitectura de la base de conocimiento, podemos afirmar que FunGramKB adopta un enfoque conceptualista, el cual facilita un tratamiento más adecuado de la interpretación semántica en sistemas multilingües del PLN.

Los esquemas conceptuales de FunGramKB desempeñan un papel primordial en la inferencia de conocimiento durante el proceso de comprensión del lenguaje. En nuestra base de conocimiento, los esquemas conceptuales se clasifican atendiendo

a dos parámetros: (i) la prototipicidad y (ii) la temporalidad. De un lado, los esquemas conceptuales almacenan conocimiento prototípico (i.e. protoestructuras), o bien pueden servir para describir una instancia de una entidad o un evento (i.e. bioestructuras). Por ejemplo, la descripción del significado de 'orden judicial' implica describir la protoestructura del concepto al que va asignada la unidad léxica; en cambio, si deseamos proporcionar información sobre Eurojust[5], necesitamos hacerlo a través de una bioestructura. Igualmente, podemos presentar el conocimiento atemporalmente (i.e. microestructuras), o inserto en un paradigma temporal (i.e. macroestructuras). Por ejemplo, la descripción del procedimiento de tramitación de una orden judicial requiere una macroestructura, mientras que una microestructura es suficiente para describir las funciones del presidente de Eurojust. Cuando combinamos estos dos parámetros, obtenemos la tipología de esquemas conceptuales mostrada en la Tabla 1:

Tabla 1. La tipología de los esquemas conceptuales en FunGramKB.

		TEMPORALIDAD	
		-	+
PROTOTIPICIDAD	+	Protomicroestructura (postulado de significado)	Protomacroestructura (guión)
	-	Biomicroestructura (retrato)	Biomacroestructura (historia)

En FunGramKB, todo este conocimiento interactúa dinámicamente durante el proceso de comprensión textual, lo cual es posible gracias a que los esquemas almacenados en cualquiera de los módulos del nivel conceptual están formalizados a través del mismo lenguaje de interfaz, i.e. COREL—COnceptual Representation Language (Periñán-Pascual & Mairal-Usón, 2010). A modo de ilustración, presentamos el postulado de significado de $WARRANT_00 (orden judicial) y su equivalente al español en (1).

(1) $+((e_1: +BE\_00\ (x_1: \$WARRANT\_00)_{Theme}\ (x_2: +DOCUMENT\_00)_{Referent})\ (e_2: +BE\_01\ (x_2)_{Theme}\ (x_3: +LEGAL\_00)_{Attribute}))$
$+(e_3: +WRITE\_00\ (x_4: +JUDGE\_00)_{Theme}\ (x_1)_{Referent}\ (f_1: (e_4: +DO\_00\ (x_5: +POLICE\_00)_{Theme}\ (x_6)_{Referent}))_{Goal})$
Una orden judicial es un documento legal.
Es emitida por un juez para permitir a la policía realizar una determinada acción.

Debido al tema que nos ocupa en este artículo, el resto de este apartado lo dedicamos a la descripción de los elementos, relaciones y propiedades de las unidades conceptuales que configuran la Ontología Nuclear, y a sus similitudes y diferencias con los conceptos de las Ontologías Satélites.

1.2. Ontología Nuclear

La Ontología Nuclear distingue tres niveles conceptuales, cada uno de los cuales está constituido por conceptos de diferente naturaleza: metaconceptos, conceptos básicos y conceptos terminales (Periñán-Pascual & Arcas-Túnez, 2010). Esta estructuración ontológica está motivada por la necesidad de poseer un nivel nuclear del conocimiento (i.e. conceptos básicos) que sirva como pivote entre aquellas categorías universales que faciliten la interoperabilidad ontológica (i.e. metaconceptos) y aquellos conceptos que puedan proporcionar una aplicabilidad inmediata (i.e. conceptos terminales). Por consiguiente, el principal objetivo de los metaconceptos (p.ej. #COMMUNICATION, #COGNITION o #MOTION) es cubrir todas las categorías cognitivas más generales que contribuyan a la estandarización y uniformidad del conocimiento semántico en el momento de integrar e intercambiar información con otras ontologías. En cambio, los conceptos básicos (p.ej. +BUILD_00, +COLD_00 o +WINDOW_00) se utilizan como componentes que ayudan a construir los postulados de significado de otros conceptos básicos y de los terminales. Por último, los conceptos terminales (p.ej. $AMAZE_00, $SUBLIMINAL_00 o $WIDOWER_00) son aquellas unidades ontológicas que carecen de potencial definitorio para formar parte en los postulados de significado de FunGramKB.

Por otra parte, la subsunción (IS-A) es la única relación taxonómica válida en el modelo ontológico de FunGramKB, lo cual implica que el nivel superior se modele en tres subontologías, donde los metaconceptos #ENTITY, #EVENT y #QUALITY sirven para organizar los nombres, verbos y adjetivos respectivamente. A pesar de esta restricción en la relación taxonómica, la Ontología Nuclear consigue un alto grado de conectividad entre sus unidades ontológicas por medio de los componentes conceptuales que comparten sus postulados de significado.

Finalmente, los conceptos de la Ontología Nuclear están provistos de una serie de propiedades semánticas, tales como el marco temático y el postulado de significado. Un marco temático es un constructo conceptual que expresa el número y tipo de participantes que intervienen en la situación cognitiva prototípica descrita por un evento o una cualidad. Estos participantes no siempre estarán instanciados lingüísticamente, pero en cambio siempre desempeñan un papel fundamental en el espacio cognitivo, de tal manera que es imposible entender el concepto sin tenerlos en cuenta. A modo de ilustración, presentamos el marco temático de +FLOAT_00:

(2) $(x_1)_{\text{Theme}}$ $(x_2$: +LIQUID_00$)_{\text{Location}}$

Así, el marco temático (2) describe un escenario cognitivo prototípico en el que 'una entidad (x_1) se encuentra en un líquido (x_2)'. Por otra parte, un postulado de significado es un conjunto de una o más predicaciones conectadas lógicamente (e_1, e_2... e_n) que representan los rasgos genéricos de los conceptos. Por ejemplo, presentamos en (3) el postulado de significado de +FLOAT_00:

(3) $+(e_1$: +LIE_00 $(x_1)_{Theme}$ $(x_2)_{Location}$ $(f_1$: $(e_2$: n +SINK_00 $(x_3)_{Agent}$ $(x_1)_{Theme}$ $(x_2)_{Location}$ $(x_4)_{Origin}$ $(x_5)_{Goal}))_{Result})$

Algo se encuentra sobre un líquido sin que se hunda.

Ya que los participantes del marco temático son necesarios cognitivamente, estos deben estar presentes en el postulado de significado de su correspondiente concepto. Más concretamente, cada participante del marco temático debe estar coindizado con un participante del postulado de significado. Por consiguiente, los marcos temáticos se integran plenamente con los postulados de significado, convirtiéndose así en las dos caras de una misma moneda.

1.3. Ontología Satélite

Las Ontologías Satélites de FunGramKB se construyen como módulos conceptuales específicos de un dominio especializado, los cuales deben estar conectados a la Ontología Nuclear, ya que:

> "specialized knowledge (ultimately) is based on and derived from everyday knowledge, for the obvious reasons that it can only be acquired on the basis of what people already know" (van Dijk, 2003: 27).

A imagen y semejanza del modelo conceptual de la Ontología Nuclear, las Ontologías Satélites también se estructuran jerárquicamente a través de la relación de subsunción. Igualmente, las propiedades conceptuales en forma de marcos temáticos y postulados de significado también se utilizan como esquemas independientes de la lengua que permiten la descripción formal del significado léxico desde el enfoque de la semántica profunda.

No obstante, las Ontologías Satélites difieren de la Ontología Nuclear en cuatro aspectos. En primer lugar, las Ontologías Satélites solo se estructuran en dos niveles conceptuales (i.e. básico y terminal), ya que estas utilizan las mismas dimensiones metaconceptuales de la Ontología Nuclear. En definitiva, solo puede haber un único espacio conceptual en el que se distribuyan tanto los conceptos de propósito general como los especializados. En segundo lugar, el inventario de conceptos básicos de cada Ontología Satélite es específico de un dominio, por lo cual este inventario léxico no coincide ni con el de la Ontología Nuclear ni con los de otras Ontologías Satélites.

En tercer lugar, el superordinado de un concepto perteneciente a una determinada Ontología Satélite puede tomar la forma de un concepto básico de la misma Ontología Satélite (i.e. subsunción intramodular), o bien un concepto básico o terminal en la Ontología Nuclear (i.e. subsunción intermodular). Finalmente, la Ontología Satélite requiere la presencia de lo que denominamos 'concepto espejo', i.e. un concepto que se sitúa inicialmente en la Ontología Nuclear, el cual tiene asignado allí un postulado de significado sobre el conocimiento del sentido común, pero al que queremos incorporar predicaciones sobre el conocimiento del dominio especializado desde la Ontología Satélite.

2. Metodología para la construcción de una ontología satélite

Desde el punto de vista de la ingeniería del conocimiento, el desarrollo de cualquier Ontología Satélite en FunGramKB sigue una metodología de trabajo dividida en las siguientes fases:

(i) Adquisición. La herramienta *FunGramKB Term Extractor* (FTE), la cual pertenece al entorno de desarrollo online *FunGramKB Suite*, permite no solo introducir un corpus a partir del cual se identifican automáticamente los términos candidatos de acuerdo con su peso probabilístico sino también asistir a los lingüistas en el proceso de evaluación de los términos relevantes.

(ii) Conceptualización. A través de FTE, se agrupan los términos resultantes de la fase (i) en unidades conceptuales, donde un término puede estar vinculado a varios conceptos (i.e. polisemia) y muchos términos pueden ser asignados a un mismo concepto (i.e. sinonimia). Esta fase requerirá además la descripción del significado de los nuevos conceptos, ya que serán los postulados de significado los que actúen como organizadores ontológicos en la fase (iv).

(iii) Selección. Se elabora automáticamente el inventario de conceptos básicos tras el procesamiento (i.e. tokenización y lematización, principalmente) de las descripciones semánticas que se asignaron en la fase (ii). Más concretamente, el criterio de pertenencia al nivel conceptual básico se basa en el índice de frecuencia de las palabras constituyentes de los textos definitorios, una vez que éstos han sido desprovistos de sus palabras funcionales (i.e. artículos, preposiciones, etc.).

(iv) Jerarquización. Tras distribuir los conceptos básicos y terminales entre las diversas dimensiones metaconceptuales de FunGramKB, éstos se organizan jerárquicamente de acuerdo con la relación de subsunción y respetando siempre los principios de similitud, especificidad y oposición como compromisos ontológicos (Periñán-Pascual & Arcas-Túnez, 2010). Para ello, es preciso construir el postulado de significado de cada uno de esos conceptos de la Ontología Satélite a partir de los textos definitorios de la fase (ii), donde se pondrá especial énfasis tanto en la granularidad como en la exactitud de la

información presentada en las representaciones formales en COREL. Con el fin de integrar los módulos ontológicos, no debemos olvidarnos de conectar los conceptos raíz en el nivel básico de la Ontología Satélite con algún concepto básico o terminal de la Ontología Nuclear.

(v) Refinamiento. Finalmente, los conceptos básicos que no desempeñen su papel ontológico de manera productiva son convertidos en conceptos terminales. Más concretamente, el umbral de productividad de los conceptos básicos resultantes de la fase anterior se establecerá automáticamente a partir de su expresividad semántica, o "contenido de información" (IC). Existen diversas medidas que nos permiten cuantificar el IC extrínseca o intrínsecamente. Desde un enfoque extrínseco, el IC de un concepto se obtiene combinando el conocimiento de la estructura jerárquica de la ontología con la estadística proveniente de un determinado corpus. Los primeros modelos sobre IC (Jiang & Conrath, 1997; Lin, 1998) adoptaron este enfoque, todos los cuales se basaron de alguna forma en el negativo del logaritmo de la verosimilitud propuesto por Resnik (1995), i.e. $-\log p(c)$, donde c es un concepto en WordNet y $p(c)$ es la probabilidad de encontrar una instancia de c en un determinado corpus. En estos casos, el IC es inversamente proporcional a la probabilidad de un concepto en el corpus: cuanto más probable es un concepto, menor será su informatividad. Ya que el estatus de un concepto básico en FunGramKB viene determinado por su uso en los postulados de significado de la ontología, es preciso adoptar un enfoque intrínseco, donde el IC se obtiene exclusivamente a partir del conocimiento ontológico, sin dependencia alguna con un corpus externo. Por ejemplo, Seco, Veale & Hayes (2004) diseñaron una medida que se apoya únicamente en la relación de subsunción de la ontología, por la cual el IC es inversamente proporcional al número de hipónimos del concepto en cuestión. Sin embargo, nosotros debemos utilizar una medida que no solo tenga en cuenta la relación taxonómica, sino también otras relaciones conceptuales más complejas. Este es el caso de la métrica de Seddiqui & Aono (2010), donde una buena parte de la información intrínseca de un determinado concepto se haya en todas las relaciones ontológicas que se puedan establecer con dicho concepto. En definitiva, y en la línea de este último modelo, el umbral de productividad de un concepto básico c en FunGramKB vendrá definido por una métrica que tenga en consideración dos factores determinantes: el número de conceptos subordinados a c y el número de predicaciones en que aparece c en otros postulados de significado.

Una vez concluida la Ontología Satélite, el trabajo se traslada al Lexicón, ya que debemos desarrollar las entradas léxicas de los términos vinculados a los conceptos de esa Ontología Satélite. De esta forma, las unidades léxicas que pertenecen al conocimiento especializado de un determinado dominio son tratadas

computacionalmente desde dos perspectivas diferentes pero complementarias. Por una parte, la dimensión léxico-sintáctica tiene lugar en el Lexicón y el Gramaticón, donde se emplean los modelos funcionales de la Gramática del Papel y la Referencia y el Modelo Léxico-Construccional para describir el comportamiento sintáctico de las unidades léxicas a través de la activación desde el Lexicón de los esquemas construccionales almacenados en el Gramaticón. Por otra parte, la dimensión semántico-conceptual tiene lugar principalmente en la Ontología Satélite, donde no solo se descubren las relaciones semánticas que se establecen entre los constituyentes de las unidades fraseológicas sino también se define el sistema conceptual del dominio a través de las relaciones establecidas entre las entidades. De hecho, la dimensión ontológica configura los cimientos de toda la base de conocimiento, incluso hasta el punto de modelar las teorías lingüísticas anteriores desde una orientación más conceptualista.

En el siguiente apartado, describimos con más detalle las fases (i) y (ii), las cuales requieren el uso de FTE. Nuestro objetivo será el desarrollo de GLOBALCRIMETERM, una Ontología Satélite sobre el derecho penal en el dominio del terrorismo y el crimen organizado, aunque la metodología descrita sirve igualmente para la construcción de cualquier otra Ontología Satélite.

3. El extractor terminológico de FunGramKB

3.1. Adquisición terminológica

El proceso semiautomático de extracción terminológica para la construcción de GLOBALCRIMETERM se dividió en cinco fases: (i) elaboración de los filtros, (ii) construcción e indización del corpus, (iii) extracción de n-gramas y cálculo estadístico, (iv) identificación de los términos y (v) validación del corpus. Se trata de un proceso semiautomático porque las fases (iii) y (v) se realizaron de manera automática, las fases (i) y (iv) de manera asistida, y en la fase (ii) se empleó un método híbrido. En las siguientes secciones, describimos con más detalle cada una de estas fases.

3.1.1. Elaboración de los filtros

Ya que GLOBALCRIMETERM es el primer corpus que se generó con FTE, fue preciso una fase inicial de preprocesamiento donde se confeccionaron las listas de filtrado a modo de *stopwords*. En realidad, no se trata de listas de palabras sino de raíces, para lo cual se aplicó el analizador morfológico Snowball basado en el algoritmo de Porter (1980) sobre los inventarios léxicos que se indican en la Tabla 2[6].

Tabla 2. Los filtros de procesamiento en la extracción terminológica.

		Inglés	Español
Funcional	Palabras funcionales de la lengua[7]	267	365
Básico	Palabras del nivel inicial e intermedio para el estudio de la lengua como idioma extranjero[8]	6.626	1.936
Avanzado	Palabras del nivel avanzado para el estudio de la lengua como idioma extranjero[9]	20.399	35.483

Ya que estos inventarios léxicos no son específicos del dominio de la criminología, podremos reutilizar las mismas listas de filtrado para otras Ontologías Satélites, por lo cual esta fase de preprocesamiento no sería necesaria en posteriores proyectos.

3.1.2. Construcción e indización del corpus

Tras la fase preliminar, el proceso de adquisición terminológica continuó con la construcción manual del corpus y su indización automática. En primer lugar, la creación del corpus implicó la especificación de información sobre descriptores como (i) el nombre [i.e. GLOBALCRIMETERM], (ii) el tipo de contenido [i.e. *Una colección representativa de textos (actas, acuerdos, informes, sentencias, libros académicos y artículos, entre otros muchos), emitidos por instituciones nacionales e internacionales, que tratan sobre el terrorismo y el crimen organizado transnacional*] y (iii) el dominio temático [i.e. *Law*][10], mientras que para cada documento que configura el corpus se identificó (iv) el título [e.g. *ETRep 2010 trend report*], (v) la descripción [e.g. *EU Terrorism situation and trend report 2010*] y (vi) el idioma [e.g. English]. Además de almacenar la versión original de cada documento, los documentos del corpus se indizaron a través de la tecnología Lucene. Net (Hatcher, Gospodnetic & McCandless, 2010), una librería de código abierto utilizada en los sistemas de recuperación de información que permite tanto indizar una extensa colección de documentos como buscar cadenas de texto en dicha colección de manera rápida. En este proceso de indización, el análisis lingüístico del corpus fue prácticamente inexistente, limitándose a un simple proceso de tokenización y a la extracción de las raíces de estos *tokens* a través del analizador morfológico Snowball. De esta forma, GLOBALCRIMETERM consta finalmente de 621 documentos y 5.698.754 *tokens*.

3.1.3. Extracción de n-gramas y cálculo estadístico

Una vez almacenado nuestro repositorio documental, se extrajeron todos los n-gramas[11] del corpus (en nuestro caso, unigramas, bigramas y trigramas) y se calculó su peso estadístico. Las investigaciones en extracción terminológica automática, muchas de las cuales se han centrado en el descubrimiento de unidades multiléxicas (por ejmplo: compuestos nominales o locuciones idiomáticas) y colocaciones,

suelen adoptar un enfoque estadístico, lingüístico o híbrido. Por una parte, existe un amplio repertorio de medidas estadísticas de asociación léxica, las cuales suelen centrar el foco en dos aspectos diferentes: (i) la significancia de la asociación, p.ej. puntuación T (Church, Gale, Hanks & Hindle, 1991) o logaritmo de la razón de verosimilitudes (Dunning, 1994) y (ii) el grado de asociación, p.ej. información mutua (Kenneth & Hanks, 1989) o cociente de probabilidades (Blaheta & Johnson, 2001)[12]. Por otra parte, podemos adoptar un enfoque que tenga en consideración las propiedades lingüísticas de los términos y de su contexto, tal y como ocurre en el programa LEXTER (Bourigault, 1992). Finalmente, se han diseñado medidas que permiten combinar los enfoques estadístico y lingüístico, por ejemplo, valor-C/valor-NC (Frantzi & Ananiadou, 1997; Frantzi, Ananiadou & Hideki, 2000). Aunque los tres enfoques anteriores se han apoyado tradicionalmente en el procesamiento de un corpus que represente el dominio objeto de estudio, también se han aplicado técnicas que permiten realizar un análisis contrastivo entre córpora, i.e. entre un corpus de referencia (corpus no técnico) y un corpus de análisis (corpus específico de un dominio); el peso contrastivo (Basili, Moschitti, Pazienza & Zanzotto, 2001) o la medida de terminologización (Wong, Liu & Bennamoun, 2007) sirven para ilustrar este último enfoque.

Nuestro extractor terminológico utiliza una adaptación de la medida *tf-idf* (Salton & Buckley, 1988) sobre el repositorio documental. Esta medida, la cual se aplica con frecuencia en la recuperación de información, nos permite ponderar la relevancia de los n-gramas con respecto a toda una colección de textos. En primer lugar, se identificaron los unigramas y se calculó su estadística; posteriormente, se hizo lo mismo con los bigramas y trigramas de forma paralela. El reconocimiento automático de bigramas y trigramas como candidatos terminológicos no se basó en patrones sintagmáticos predefinidos, pero sí se tuvo en cuenta la distribución de los términos funcionales dentro de estas unidades complejas. Así, en el caso de los bigramas, se eliminaron aquellos que contenían la raíz de una palabra funcional. En el caso de los trigramas, se eliminaron aquellos que contenían la raíz de una palabra funcional al principio o al final del trigrama.

A continuación, describimos brevemente nuestra adaptación de la medida *tf-idf* para determinar el peso y la relevancia de los n-gramas en el corpus:

(i) Calculamos la frecuencia del n-grama relativa a todo el corpus (*tf*)[13], lo cual se implementó como la raíz cuadrada de *f*, i.e. la frecuencia total del n-grama[14].

(ii) Calculamos la frecuencia inversa del documento (*idf*), implementada como:

$$log\ (numDocs\ /\ (docFreq + 1)) + 1$$

donde *numDocs* es el número total de documentos que configuran el corpus, y *docFreq* es el número de documentos que contienen el n-grama.

(iii) Calculamos el producto de la frecuencia del término y la frecuencia inversa del documento (*tf* idf*), redondeando el resultado a dos decimales.

3.1.4. Identificación de los términos

Al ordenar los n-gramas por su índice *tf-idf*, FTE nos permitió descubrir más fácilmente los términos relevantes para nuestro dominio de estudio. No obstante, fue preciso realizar una tarea de 'depuración terminológica' a través de una interfaz similar a la de la Figura 1.

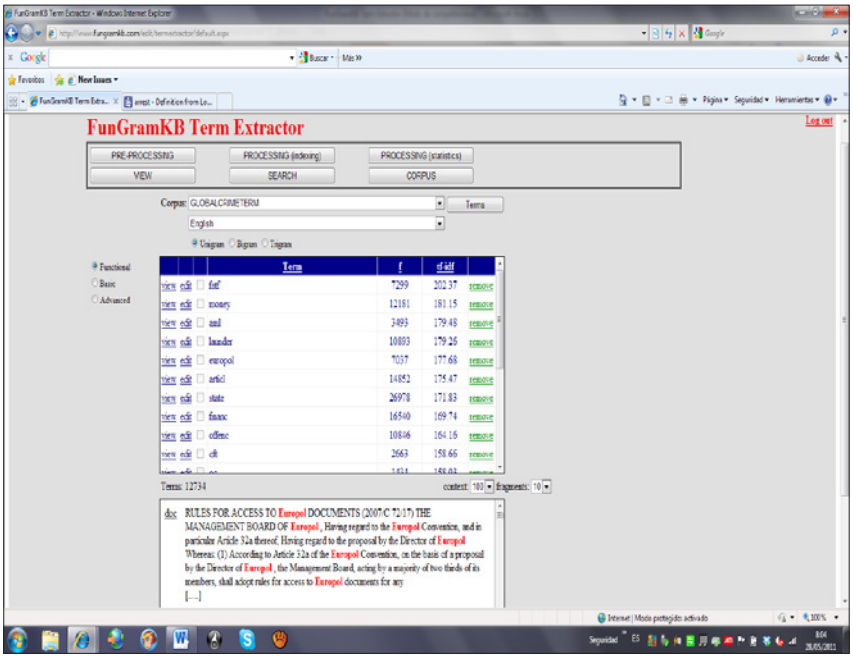

Figura 1. El visualizador de n-gramas.

Desde este editor, el investigador, ya sea lingüista o ingeniero del conocimiento, puede realizar las siguientes tareas:

(i) Extraer del corpus aquellos fragmentos más relevantes donde aparece el n-grama en cuestión, para lo cual podemos indicar la longitud y número de estos fragmentos.
(ii) Eliminar temporalmente un n-grama por no servir como término relevante al dominio del corpus.
(iii) Crear el concepto asociado al término cuyo n-grama hemos seleccionado. A esta tarea la denominamos 'conceptualización terminológica', y será descrita con mayor detalle en el apartado 3.2.

En la mayoría de sistemas de extracción automática de términos, solo los candidatos que reciben una puntuación superior a un determinado umbral son enviados a validación. En FTE, el investigador tiene a su disposición todos los n-gramas a los que se le aplicó un índice *tf-idf* en la fase anterior. De esta forma, podemos construir manualmente un inventario terminológico de mayor precisión y cobertura para nuestra Ontología Satélite[15].

El proceso de depuración terminológica consiste básicamente en la eliminación de los candidatos falsos a partir de los n-gramas recuperados. Aquí diferenciamos cinco niveles de trabajo, resultantes de combinar el tipo de n-grama (i.e. unigrama, bigrama y trigrama) y, en el caso de los unigramas, el tipo de filtrado (i.e. funcional, básico y avanzado). La Tabla 3 muestra la secuenciación de estos niveles de trabajo:

Tabla 3. Los niveles en la depuración terminológica.

Nivel	Objeto de la tarea
A	Trigramas
B	Bigramas
C	Unigramas con filtrado avanzado
D	Unigramas con filtrado básico
E	Unigramas con filtrado funcional

En realidad, la tarea de depuración terminológica se basa en un método en cascada (Jurafsky & Martin, 2009), donde es preciso que termine el filtrado de n-gramas de un nivel antes de que empiece el siguiente. En nuestro caso, este método solo puede ser aplicado de manera descendente, permitiendo recorrer los n-gramas desde el más específico hasta el más general en cuanto a su contenido semántico. Así, los trigramas (nivel A) son potencialmente más restrictivos en cuanto a su contenido semántico que los unigramas (niveles C-E) y el filtrado avanzado (nivel C) lo es más que el filtrado funcional (nivel E). En sentido estricto, la depuración terminológica conlleva simplemente la decisión de eliminar el n-grama, o dejarlo como término relevante para su conceptualización. Si un n-grama resulta ser un candidato falso, podemos eliminar dicho n-grama de manera parcial o total:

(i) El n-grama debe ser eliminado de manera parcial cuando alguno de los n-gramas componentes ($n - 1 > 0$) pueda servir como término relevante en niveles posteriores de depuración. Así, el trigrama *'combat(ing) money laundering'* que nos encontramos en el nivel A fue eliminado por no ser relevante para su conceptualización, pero la depuración fue parcial, porque el bigrama *'money laundering'* se convirtió a concepto en el nivel B. En otras palabras, nos encontramos un n-grama anidado relevante en un n-grama más grande no relevante.

(ii) El n-grama debe ser eliminado de manera total cuando ninguno de los n-gramas componentes ($n - 1 > 0$) pueda servir como término relevante en niveles

posteriores de depuración. Así, el trigrama *'financing of crimes'* fue eliminado de manera total desde el nivel A, depurando indirectamente los niveles inferiores de trabajo, por lo cual también se eliminó de manera automática *'financing'* y *'crimes'*, los cuales pertenecían al nivel D. En otras palabras, nos encontramos un n-grama no relevante que tampoco contiene n-gramas anidados relevantes.

Una de las cuestiones más espinosas en esta fase del procesamiento fue precisamente determinar el criterio que motivaba la depuración de los bigramas y los trigramas. En primer lugar, y de acuerdo con la perspectiva contextualista de Sinclair (1996, 1998), se identificaron los casos de colocaciones, i.e. combinaciones léxicas que co-ocurren frecuentemente, y preferencias de selección, i.e. campo semántico de dichas colocaciones. Mientras la información sobre las preferencias de selección fue registrada para su posterior conceptualización e inclusión en los marcos temáticos durante la fase de jerarquización, los colocados tuvieron que pasar una criba semántica. En segundo lugar, por tanto, las unidades fraseológicas[16] fueron categorizadas como 'transparentes' u 'opacas', dependiendo de si su significado podría o no ser inferido por la máquina a través de un proceso de composicionalidad de los postulados de significado vinculados a los conceptos formantes de la unidad fraseológica en cuestión. De esta manera, *'aggravating factors'*, *'drug dealer'*, *'protection of witenesses'* o *'terrorist group'* son unidades fraseológicas transparentes, muchas de las cuales son tratadas como casos de colocación léxica. Por otra parte, *'frozen account'*, *'financial haven'* o *'white-collar crime'* son unidades fraseológicas opacas, donde el humano aplica algún mecanismo de extensión metafórica o metonímica para su comprensión. Esta decisión tuvo una importante repercusión en la siguiente fase del modelado ontológico, donde permitimos la conceptualización en una sola unidad ontológica de cada una de las unidades fraseológicas opacas, además de cada unidad fraseológica transparente que a nivel cognitivo desempeñe la función de concepto espejo. Observamos que el estudio de las unidades fraseológicas especializadas en FunGramKB se encuadra en un marco metodológico que integra los enfoques estadístico y semántico, tal y como sugiere Corpas Pastor (2001). En otras palabras, en cada combinación léxica de co-ocurrencia extraída probabilísticamente, se considera la relación de dependencia semántica entre la base y su colocativo, donde algunas unidades fraseológicas transparentes formarán parte del Lexicón, mientras que las unidades fraseológicas opacas se vincularan a unidades conceptuales de la Ontología, donde además las preferencias de selección serán incluidas en los marco temáticos de dichos conceptos.

3.1.5. Validación del corpus

Finalmente, la validación en FTE bloquea el modo de edición del corpus, sin permitirnos extraer más términos ni introducir más documentos, además de eliminar de manera definitiva todos aquellos n-gramas que se marcaron como candidatos falsos, por lo cual se trata en definitiva de una fase de limpieza.

3.2. Conceptualización terminológica

En esta tarea, para cada término relevante, el cual puede estar formado por una o varias palabras ortográficas, creamos uno o más conceptos, los cuales podrán tomar una de las siguientes formas:

(i) un concepto de la Ontología Satélite, p.ej. +FINANCIAL_INTELLIGENCE_UNIT_00 (unidad de inteligencia financiera)[17],

(ii) un concepto de la Ontología Nuclear, p.ej. $MONEY_LAUNDERING_00 (lavado de dinero), $SMUGGLING_00 (contrabando) o $FORGERY_00 (falsificación), o

(iii) un concepto del Onomasticón, p.ej. %EUROJUST_00, %EUROPOL_00 o %OSCE_00 (*Organization for Security and Co-operation in Europe*) o %SEPBLAC_00 (Servicio Ejecutivo de la Comisión de Prevención del Blanqueo de Capitales e Infracciones Monetarias).

En el caso de (i), FTE nos ofrece la posibilidad de poder describir la información conceptual básica que permitirá dar de alta al concepto en cuestión, tal y como muestra la Figura 2. Posteriormente, el ingeniero del conocimiento podrá asignar las propiedades conceptuales necesarias (i.e. postulado de significado y marco temático) desde el editor ontológico correspondiente.

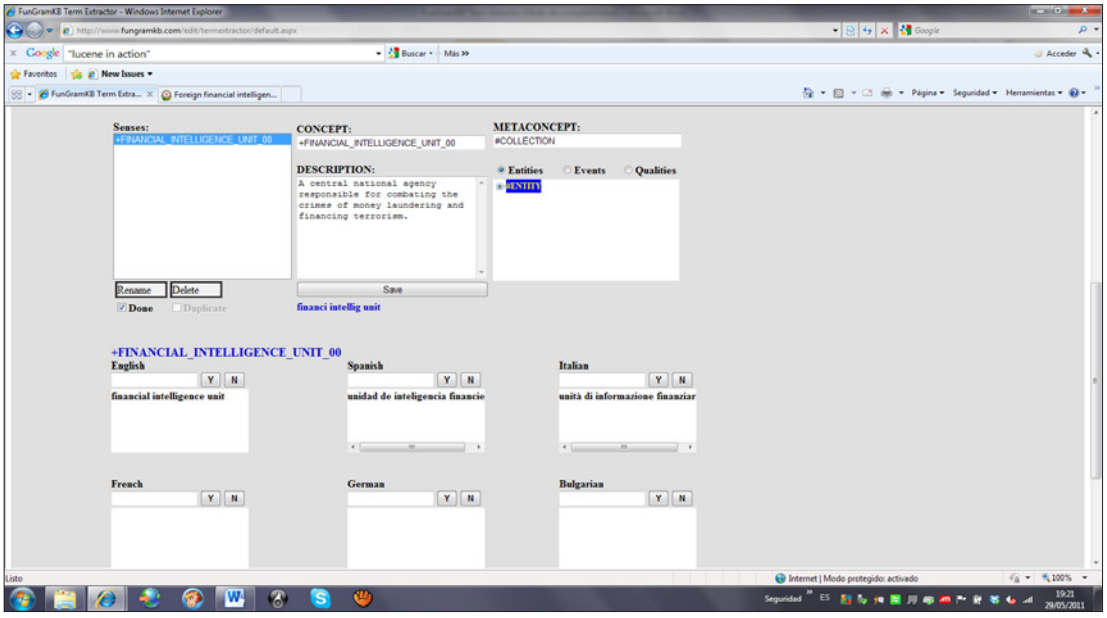

Figura 2. La interfaz de conceptualización.

En el caso de (ii), es posible que un concepto como $MONEY_LAUNDERING_00 ya esté definido en la Ontología Nuclear como 'la acción de invertir en empresas legales dinero obtenido ilegalmente', pero precise un concepto espejo en la Ontología Satélite con el fin de almacenar información correspondiente a la sanción por lavado de dinero, p.ej. los años de privación de libertad y/o la multa económica. Otras unidades léxicas como *'crime'*, *'judge'* o *'syndicate'*, entre otras muchas, también estarán vinculadas a conceptos espejos en la Ontología Satélite con el fin de poder incorporar nuestro conocimiento experto. Por esta razón, preferimos no hablar de términos o unidades fraseológicas especializadas sino más bien de unidades léxicas vinculadas al conocimiento especializado.

No solo podemos almacenar conocimiento semántico, sino también podemos incorporar el conocimiento procedimental a través de los guiones del Cognicón o las historias del Onomasticón. Por ejemplo, aunque hayamos definido qué es Eurojust en el Onomasticón, también es posible representar formalmente las diversas fases que intervienen en el procedimiento de votación para la elección del presidente de esta agencia (#ELECTING_THE_PRESIDENT_OF_EUROJUST_00), lo cual requerirá una historia como constructo cognitivo.

CONCLUSIONES

Las bases de datos digitales de contenido legal están adquiriendo dimensiones considerables, pero en su mayor parte se trata de meros contenedores de textos enfocados a la comprensión humana. En este contexto, se emplean métodos de gestión documental poco eficaces que orientan las consultas de los usuarios a las formas léxicas presentes en los textos legales, desatendiendo su contenido semántico. El PLN basado en la semántica profunda puede brindar al ámbito del derecho funcionalidades tales como la realización de consultas basadas en lenguaje natural, la extracción de información en forma de resúmenes de sentencias, la clasificación automática de los documentos de un proceso judicial, la identificación de las partes implicadas y las resoluciones adoptadas, e incluso servir de guía de razonamiento sobre los argumentos de las declaraciones de las partes. Como un primer paso hacia la construcción de dichas aplicaciones computacionales, este artículo ha descrito la metodología empleada en el desarrollo de GLOBALCRIMETERM, una Ontología Satélite incorporada a FunGramKB sobre el terrorismo y el crimen organizado dentro del dominio legal. Este trabajo se centra particularmente en las etapas de adquisición y conceptualización de los términos, las cuales requieren el uso de la herramienta FTE. La semántica profunda de esta ontología terminológica, junto con la implementación de mecanismos cognitivos que simulen la capacidad del razonamiento humano, la convierten en un recurso extraordinariamente útil para su explotación en sistemas del PLN que requieran la comprensión del lenguaje, y especialmente en lo concerniente al desarrollo de agentes inteligentes con una interfaz de comunicación persona-máquina.

Con el fin de destacar los beneficios de la metodología empleada en la construcción de las Ontologías Satélites, presentamos en primer lugar los problemas más recurrentes en la extracción terminológica a través de las herramientas informáticas incorporadas en la mayoría de los programas de gestión de córpora:

> (1) hay combinaciones muy frecuentes que no presentan un grado de estabilidad suficiente para ser consideradas colocaciones; (2) hay colocaciones muy estables cuyos colocados son palabras poco frecuentes, por lo que no aparecen en un corpus dado; (3) hay colocaciones cuyos elementos aparecen muy distanciados en el discurso, por lo que no pueden ser extraídos de forma automática; (4) la frecuencia estadística no puede dar cuenta de la prominencia cognitiva […] de algunas colocaciones muy establecidas y típicas de una lengua; (5) los programas de gestión de corpus no están diseñados para detectar colocaciones en el nivel lexemático, solo en el nivel de la palabra gráfica […]; y (6) el enfoque estadístico no dispone de instrumentos para el análisis semántico de una determinada colocación (Corpas Pastor, 2001: 100).

En nuestro proceso de adquisición terminológica, el índice *tf*-idf, el cual nos permite calcular la relevancia de los unigramas, brigramas y trigramas con respecto a todo el corpus textual, nos ayuda a aliviar los problemas (1) y (4), aunque no evita que sea preciso evaluar manualmente los términos candidatos extraídos por FTE. En cambio, los problemas (2) y (5) son solventados por completo. Por una parte, FTE permite aplicar diversos niveles de depuración terminológica (i.e. funcional, básico y avanzado) basados en la frecuencia de uso de los constituyentes léxicos fuera del dominio especializado. Por otra parte, el peso *tf-idf* se calcula a partir de las raíces de las palabras, resultando más efectivo que cuando nos centramos en la forma léxica, o incluso en el lexema, para determinar la relevancia terminológica. Aunque el problema (6) es intrínseco a cualquier método estocástico durante el proceso de adquisición terminológica, los compromisos ontológicos de FunGramKB guían la modelación del conocimiento especializado en las fases de conceptualización y jerarquización. Finalmente, el problema (3) nos marca una nueva ruta de trabajo para futuras investigaciones.

REFERENCIAS BIBLIOGRÁFICAS

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM, 26*(11), 832-843.

Basili, R., Moschitti, A., Pazienza, M. T. & Zanzotto, F. M. (2001). *A contrastive approach to term extraction*. Ponencia presentada en el 4th Terminological and Artificial Intelligence Conference. Nancy, Francia.

Benjamins, V., Contreras, J., Casanovas, P., Ayuso, M., Becue, M., Lemus, L. & Urios, C. (2004). Ontologies of professional legal knowledge as the basis for intelligent IT support for judges. *Artificial Intelligence and Law, 12*(4), 359-378.

Blaheta, D. & Johnson, M. (2001). *Unsupervised learning of multi-word verbs*. Ponencia presentada en el ACL Workshop on Collocations. Toulouse, Francia.

Bourigault, D. (1992). *Surface grammatical analysis for the extraction of terminological noun phrases*. Ponencia presentada en el 14th International Conference on Computational Linguistics. Nantes, Francia.

Breuker, J., Casanovas, P., Klein, M. C. A. & Francesconi, E. (2009). *Law, ontologies and the Semantic Web: Channelling the legal information flood*. Ámsterdam: IOS Press.

Bustos, J. M. (2001). Definición de glosarios léxicos del español: Niveles inicial e intermedio. *Enseñanza*, 19, 35-72.

Casanovas, P., Casellas, N. & Vallbé, J. (2009). An ontology-based decision support system for judges. En J. Breuker, P. Casanovas, M. Klein & E. Francesconi (Eds.), *Law, ontologies and the Semantic Web: Channelling the legal information flood* (pp. 165-175). Ámsterdam: IOS Press.

Casanovas, P., Sartor, G., Biasiotti, M. A. & Fernández-Barrera, M. (2011). Theory and methodology in legal ontology engineering: Experiences and future directions. En G. Sartor, P. Casanovas, M. A. Biasiotti & M. Fernández-Barrera (Eds.), *Approaches to legal ontologies, theories, domains, methodologies* (pp. 3-14). Berlín-Heidelberg: Springer.

Church, K., Gale, W., Hanks, P. & Hindle, D. (1991). Using statistics in lexical analysis. En U. Zernik (Ed.), *Lexical Acquisition: Exploiting on-line resources to build a lexicon* (pp. 115-64). Hillsdale: Lawrence Erlbaum.

Cimiano, P., Mädche, A., Staab, S. & Völker, J. (2009). Ontology learning. En S. Staab & R. Studer (Eds.), *Handbook of ontologies* (pp. 245-267). Berlín-Heidelberg: Springer.

Cimiano, P. & Völker, J. (2005). Text2Onto: A framework for ontology learning and data-driven change discovery. *Ponencia presentada en el 10th International Conference on Applications of Natural Language to Information Systems* (pp. 227-238). Berlín-Heidelberg: Springer.

Collin, P. (2004). *Easier English basic dictionary*. Londres: Bloomsbury.

Corpas Pastor, G. (1998). Criterios generales de clasificación del universo fraseológico de las lenguas, con ejemplos tomados del español y del inglés. En M. Alvar Ezquerra & G. Corpas Pastor (Eds.), *Diccionarios, frases, palabras* (pp. 157-187). Málaga: Universidad de Málaga.

Corpas Pastor, G. (2001). En torno al concepto de colocación. *Euskera,* 46, 89-108.

Dunning, T. (1994). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Evert, S. (2004). *The statistics of word cooccurrences: Word pairs and collocations*. Tesis doctoral, Universidad de Stuttgart, Stuttgart, Alemania.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, Mass: MIT Press.

Fillmore, C. J. (1982). Frame semantics. En Linguistic Society of Korea (Ed.), *Linguistics in the morning calm: Selected papers from SICOL-1981* (pp. 111-137). Seúl: Hanshin.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6, 222-254.

Francesconi, E. & Tiscornia, D. (2008). Building semantic resources for legislative drafting: The DALOS Project. En P. Casanovas, G. Sartor, R. Rubino & N. Casellas (Eds.), *Computable Models of the Law, Lecture Notes in Computer Science* (pp. 56-70). Berlín-Heidelberg: Springer.

Frantzi, K. & Ananiadou, S. (1997). *Automatic term recognition using contextual cues*. Ponencia presentada en el IJCAI Workshop on Multilinguality in Software Industry, AIC, Japón.

Frantzi, K., Ananiadou, S. & Hideki, M. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal of Digital Libraries, 3*(2), 115-130.

Gangemi, A., Sagri, M. & Tiscornia, D. (2003). Metadata for content description in legal information. *Proceedings of the 14th International Workshop on Database and Expert Systems Applications* (pp. 745-749). Washington, DC: IEEE Computer Society.

Gangemi, A., Sagri, M. & Tiscornia, D. (2005). A constructive framework for legal ontologies. En V. R. Benjamins, P. Casanovas, J. Breuker & A. Gangemi (Eds.), *Law and the Semantic Web: Legal ontologies, methodologies, legal information retrieval, and applications* (pp. 97-124). Berlín-Heidelberg: Springer.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Gutiérrez Cuadrado, J. (1996). *Diccionario Salamanca de la lengua española*. Madrid: Santillana-Universidad de Salamanca.

Hakimpour, F. & Geppert, A. (2001). *Ontologies: An approach to resolve semantic heterogeneity in databases*. Informe técnico, Universidad de Zúrich, Zúrich, Suiza.

Hatcher, E., Gospodnetic, O. & McCandless, M. (2010). *Lucene in action*. Greenwich: Manning.

Jiang, J. & Conrath, D. (1997). *Semantic similarity based on corpus statistics and lexical taxonomy*. Ponencia presentada en el International Conference on Research in Computational Linguistics. Taiwán.

Jurafsky, D. & Martin, J. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Nueva Jersey: Prentice Hall.

Kenneth, W. C. & Hanks, P. (1989). Word association norms, mutual information, and lexicography. En *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (pp. 76-83). Association for Computational Linguistics.

Langley, P., Laird, J. E. & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research,* 10, 141-160.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning* (pp. 296-304). Madison: Morgan Kaufmann.

Mairal-Usón, R. & Periñán-Pascual, C. (2009). The anatomy of the lexicon within the framework of an NLP knowledge base. *Revista Española de Lingüística Aplicada,* 22, 217-244.

Mairal-Usón, R. & Ruiz de Mendoza, F. (2009). Levels of description and explanation in meaning construction. En Ch. Butler & J. Martín Arista (Eds.), *Deconstructing constructions* (pp. 153-198). Ámsterdam-Filadelfia: John Benjamins.

Minsky, M. (1975). A framework for representing knowledge. En P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211-277). Nueva York: McGraw-Hill.

Ovchinnikova, E. (2012). *Integration of world knowledge for natural language understanding.* París: Atlantis Press.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. *Proceedings of the ACL 2005 Student Research Workshop* (pp. 13-18). Ann Arbor: Association of Computational Linguistics.

Periñán-Pascual, C. & Arcas-Túnez, F. (2004). Meaning postulates in a lexico-conceptual knowledge base. *Proceedings of the 15th International Workshop on Databases and Expert Systems Applications* (pp. 38-42). Los Alamitos: IEEE.

Periñán-Pascual, C. & Arcas-Túnez, F. (2005). Microconceptual-Knowledge Spreading in FunGramKB. *Proceedings of the 9th IASTED International Conference on Artificial Intelligence and Soft Computing* (pp. 239-244). Anaheim-Calgary-Zúrich: ACTA Press.

Periñán-Pascual, C. & Arcas-Túnez, F. (2007a). Deep semantics in an NLP knowledge base. *Proceedings of the 12th Conference of the Spanish Association for Artificial Intelligence* (pp. 279-288). Salamanca: Universidad de Salamanca.

Periñán-Pascual, C. & Arcas-Túnez, F. (2007b). Cognitive modules of an NLP knowledge base for language understanding. *Procesamiento del Lenguaje Natural,* 39, 197-204.

Periñán-Pascual, C. & Arcas-Túnez, F. (2010). Ontological commitments in FunGramKB. *Procesamiento del Lenguaje Natural,* 44, 27-34.

Periñán-Pascual, C. & Mairal-Usón, R. (2010). La gramática de COREL: Un lenguaje de representación conceptual. *Onomázein,* 21, 11-45.

Peters, W., Sagri, M. & Tiscornia, D. (2007). The structuring of legal knowledge in LOIS. *Artificial Intelligence and Law,* 15, 117-135.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 448-453). Montreal.

Ruiz de Mendoza, F. & Mairal-Usón, R. (2008). Levels of description and constraining factors in meaning construction: An introduction to the Lexical Constructional Model. *Folia Linguistica, 42*(2), 355-400.

Sagri, M., Tiscornia, D. & Bertagna, F. (2004). Jur-WordNet. En P. Sojka, K. Pala, P. Smrz, C. Fellbaum & P. Vossen (Eds.), *Proceedings of the Second International WordNet Conference* (pp. 305-310). Brno: Universidad de Masaryk.

Saias, J. & Quaresma, P. (2003). A methodology to create legal ontologies in a logic programming information retrieval system. En V. R. Benjamins, P. Casanovas, J. Breuker & A. Gangemi (Eds.), *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications* (pp. 185-200). Berlín-Heidelberg: Springer.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513-523.

Seco, N., Veale, T. & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. *Proceedings of the 16th European Conference on Artificial Intelligence* (pp. 1089-1090). Valencia.

Seddiqui, H. & Aono, M. (2010). Metric of intrinsic information content for measuring semantic similarity in an ontology. *Proceedings of the 7th Asia-Pacific Conference on Conceptual Modeling* (pp. 89–96). Brisbane.

Sinclair, J. M. (Ed.) (1987). *Collins COBUILD English language dictionary*. Londres: HarperCollins.

Sinclair, J. M. (1996). The search for units of meaning. *TEXTUS, 9*(1), 75-106.

Sinclair, J. M. (1998). The lexical item. En E. Weigand (Ed.), *Contrastive lexical semantics* (pp. 1-24). Ámsterdam-Filadelfia: John Benjamins.

Sun, R. (1997). Learning, action, and consciousness: A hybrid approach towards modeling consciousness. *Neural Networks, 10*(7), 1317-1331.

Sun, R. (1999). Accounting for the computational basis of consciousness: A connectionist approach. *Consciousness and Cognition*, 8, 529-565.

Sun, R., Merrill, E. & Peterson, T. (2001). From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science, 25*(2), 203-244.

Tiscornia, D. (2006). *The LOIS project: Lexical ontologies for legal information sharing.* Ponencia presentada en el 5th Legislative XML Workshop, Florencia, Italia.

Valente, A. (1995). *Legal knowledge engineering: A modeling approach.* Ámsterdam: IOS Press.

Valente, A. (2005). Types and roles of legal ontologies. En V. R. Benjamins, P. Casanovas, J. Breuker & A. Gangemi (Eds.), *Law and the Semantic Web: Legal ontologies, methodologies, legal information retrieval, and applications* (pp. 65-76). Berlín-Heidelberg: Springer.

Valente, A. & Breuker, J. (1994). Ontologies: The missing link between legal theory

and AI & Law. En A. Soeteman (Ed.), *Legal knowledge based systems JURIX 94: The foundation for legal knowledge systems* (pp. 138-150). Lelystad: Koninklijke Vermande.

van Dijk, T. A. (2003). Specialized discourse and knowledge. A case study of the discourse of modern genetics. *Cadernos de Estudos Lingüísticos,* 44, 21-55.

Van Kralingen, R. W. (1995). *Frame-based conceptual models of statute law*. La Haya: Kluwer Law International.

Van Kralingen, R. W., Visser, P. R. S., Bench Capon, T. J. M. & Van den Herik, H. J. (1999). A principled approach to developing legal knowledge systems. *International Journal of Human-Computer Studies,* 54, 1127-1154.

Van-Valin, R. (2005). *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press.

Van-Valin, R. & LaPolla, R. (1997). *Syntax: Structure, meaning, and function*. Cambridge: Cambridge University Press.

Velardi, P., Navigli, R., Cuchiarelli, A. & Neri, F. (2005). Evaluation of Ontolearn, a methodology for automatic population of domain ontologies. En P. Buitelaar, P. Cimiano & B. Magnini (Eds.), *Ontology learning from text: Methods, evaluation and applications* (pp. 92-106). Ámsterdam: IOS Press.

Velardi, P., Pazienza, M. T. & Fasolo, M. (1991). How to encode semantic knowledge: A method for meaning representation and computer-aided acquisition. *Computational Linguistics, 17*(2), 153-170.

Venturi, G., Lenci, A., Montemagni, S., Vecchi, E. M., Sagri, M. T., Tiscornia, D. & Agnoloni, T. (2009). Towards a FrameNet resource for the legal domain. En N. Casellas, E. Francesconi, R. Hoekstra & S. Montemagni (Eds.), *Proceedings of the 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques: 2nd Workshop on Semantic Processing of Legal Text* (pp. 67-76). Barcelona.

Visser, P. R. S. (1995). *Knowledge specification for multiple legal tasks. A case study of the interaction problem in the legal domain*. Tesis doctoral, Universidad de Leiden, Leiden, Holanda.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities, 32*(2-3), 73-89.

Wong, W., Liu, W. & Bennamoun, M. (2007). Determining termhood for learning domain ontologies using domain prevalence and tendency. *Proceedings of the 6th Australasian Conference on Data Mining* (pp. 47-54). Gold Coast, Australia.

NOTAS

1. Un *synset* representa un conjunto de lemas de la misma categoría gramatical que pueden ser intercambiados en un determinado contexto.
2. La dimensión multidimensional de Jur-WordNet se explota en el proyecto LOIS—Legal Ontologies for Knowledge Sharing (Tiscornia, 2006; Peters, Sagri & Tiscornia, 2007).
3. www.fungramkb.com
4. La herencia múltiple no monotónica permite que un concepto tenga asignado más de un superordinado y que la información genérica de los superordinados pueda ser rebatida por la más específica de los conceptos subordinados. Periñán-Pascual & Arcas-Túnez (2010) describieron el tratamiento de este tipo de herencia en el modelo ontológico de FunGramKB.
5. Eurojust (www.eurojust.europa.eu) es una agencia de la Unión Europea cuyo objetivo es reforzar la cooperación judicial entre los Estados miembros.
6. Los números representan las raíces obtenidas tras el procesamiento morfológico de los inventarios léxicos.
7. Junto con palabras como artículos, pronombres, preposiciones y conjunciones, también se incluyeron contracciones (por ejemplo: *'s, 've*), verbos auxiliares (por ejemplo: *be, do, have*, *get*), caracteres numéricos arábigos y romanos, y algunas abreviaturas muy usuales en documentación (e.g., i.e., c.f., etc, et al…).
8. En el caso del inglés, se utilizó el léxico del *Easier English Basic Dictionary* (Collin, 2004), y para el español el inventario léxico se obtuvo de Bustos (2001). En ambos casos, se incluyeron además los alomorfos de los verbos irregulares pertinentes al nivel.
9. En el caso del inglés, se tomó el léxico del *Collins COBUILD English Language Dictionary* (Sinclair, 1987), y para el español se consideró el *Diccionario Salamanca* (Gutiérrez Cuadrado, 1996). En ambos casos, se incluyeron además los alomorfos de los verbos irregulares pertinentes al nivel.
10. Aquí empleamos el mismo inventario de dominios que el utilizado en las entradas léxicas del Lexicón de FunGramKB.
11. En nuestro caso, un n-grama es una cadena de texto formada por *n* raíces correspondientes a palabras contiguas en el texto, ya sean separadas por un espacio en blanco o por un signo de puntuación, donde $0 < n \leq 3$.
12. Véanse Evert (2004) y Pecina (2005) para una descripción, comparativa y evaluación de las medidas estadísticas de asociación léxica.
13. Tradicionalmente, *tf* se calcula a partir de la frecuencia del término dentro de un documento.
14. Si $f < 3$, entonces se eliminaba el n-grama.
15. No obstante, con el fin de agilizar el proceso de depuración, tenemos la posibilidad de marcar manualmente un determinado valor *td-idf* a partir del cual se eliminen automáticamente todos los n-gramas del mismo tipo con valores inferiores.
16. Utilizamos el término 'unidad fraseológica' para referirnos a cualquier combinación léxica cuyos rasgos distintivos son la polilexicalidad, la alta frecuencia de aparición, la institucionalización, la estabilidad, la idiomaticidad y la variación potencial (Corpas Pastor, 1998).
17. Este concepto se definiría como 'la agencia nacional que se encarga de combatir los delitos de lavado de dinero y financiamiento del terrorismo'.