LA JERARQUIZACIÓN COGNITIVA DE LA ENTIDAD +CYBERCRIME_00 EN FUNGRAMKB

MARÍA DE GRACIA CARRIÓN DELGADO
Universidad Nacional de Educación a Distancia
AMALIA MARÍN RUBIALES
Universidad de Córdoba

RESUMEN

FunGramKB es una base de conocimiento léxico-conceptual multipropósito y multilingüe diseñada para su aplicación en diversas tareas de Procesamiento del Lenguaje Natural (PLN) tales como la traducción automática o el razonamiento artificial y en varias lenguas[1] (Periñán y Arca,s 2004; Mairal y Periñán, 2009; Periñán y Mairal, 2010). Su estructura modular refleja tres niveles de conocimiento —léxico, gramatical y ontológico— que, aunque independientes, están relacionados entre sí a través del módulo conceptual, que es compartido por todas las lenguas integradas en la base de conocimiento. Por tanto, la ontología la componen dos módulos: un módulo de propósito general, es decir, la ontología nuclear, y varios módulos terminológicos de dominios específicos, esto es, las ontologías satélite. De hecho, la ontología nuclear sirve de eje angular de toda la base de conocimiento a la vez que denota conocimiento del sentido común; mientras que las ontologías terminológicas enriquecen a la ontología nuclear a través del modelado de conocimiento especializado. En la presente contribución nos centramos en una ontología satélite del ámbito legal vinculada a la ontología nuclear de FunGramKB[2], y más concretamente en el ámbito del crimen organizado, donde analizamos cómo se desarrolla la jerarquización cognitiva de los delitos típicos de la criminalidad informática asociados a la entidad +CYBERCRIME_00. Para la elaboración de la jerarquización presentamos, por un lado, la metodología COHERENT, base de la ontología nuclear y marco teórico de referencia en el desarrollo de las ontologías satélite vinculadas a ella y, por otro, el lenguaje de representación conceptual COREL, que sirve de base angular a los ingenieros del conocimiento en el desarrollo de la jerarquía conceptual del dominio objeto de estudio.

PALABRAS CLAVE: FunGramKB, base de conocimiento, Procesamiento del Lenguaje Natural (PLN), ontología nuclear, ontología satélite, jerarquización cognitiva, COHERENT, COREL

ABSTRACT

FunGramKB is a multilingual and multipurpose lexico-conceptual knowledge-base designed for its use in various tasks in Natural Language Processing (NLP) such as machine translation or artificial reasoning and in

several languages[1] *(Periñán and Arcas, 2004; Mairal and Periñán, 2009, Mairal and Periñán, 2010). Its modular structure reflects three levels of knowledge—lexical, grammatical and ontological— which, though independent, are interrelated through the conceptual module, which is shared by all the languages integrated in the knowledge base. Therefore, the ontology comprises two modules: a general purpose one, i.e. the core ontology, and various modules of terminological specific domains, that is, the satellite ontologies. In fact, the core ontology serves as the angular axis of the knowledge base while it denotes the common sense knowledge, whereas the terminological ontologies enrich the core ontology through the modeling of expert knowledge. In this contribution we focus on a legal satellite ontology linked to the FunGramKB*[2] *core ontology: in the area of organized crime, where we analyze how to develop the cognitive hierarchy of the typical computer-related crimes associated with the entity +CYBERCRIME_00. For the preparation of the present hierarchy, we introduce, on the one hand, the COHERENT methodology,base of the core ontology and theoretical framework in the development of satellite ontologies linked to it and, on the other, the conceptual representation language COREL, which serves as angular base for the knowledge engineers in the development of the conceptual hierarchy of the domain under scrutiniy.*

KEYWORDS: FunGramKB, knowledge base, Natural Language Processing (NLP),core ontology, satellite ontology, cognitive hierarchy, COHERENT, COREL

1. INTRODUCCIÓN

FunGramKB es una base de conocimiento léxico-conceptual diseñada para la realización de tareas de procesamiento del lenguaje natural. Es multipropósito y multilingüe porque puede ser reutilizada en varias lenguas en tareas de recuperación y extracción de información o traducción automática, entre otras.

En *FunGramKB* aparecen representados los tres niveles principales de conocimiento (léxico, gramatical y conceptual) y, aunque independientes, están interrelacionados entre sí (Periñán y Arcas, 2011: 2-3).

Así, los módulos léxico y gramatical son dependientes de cada lengua, mientras que el módulo conceptual es compartido por todas las lenguas integradas en la base de conocimiento (Figura 1).

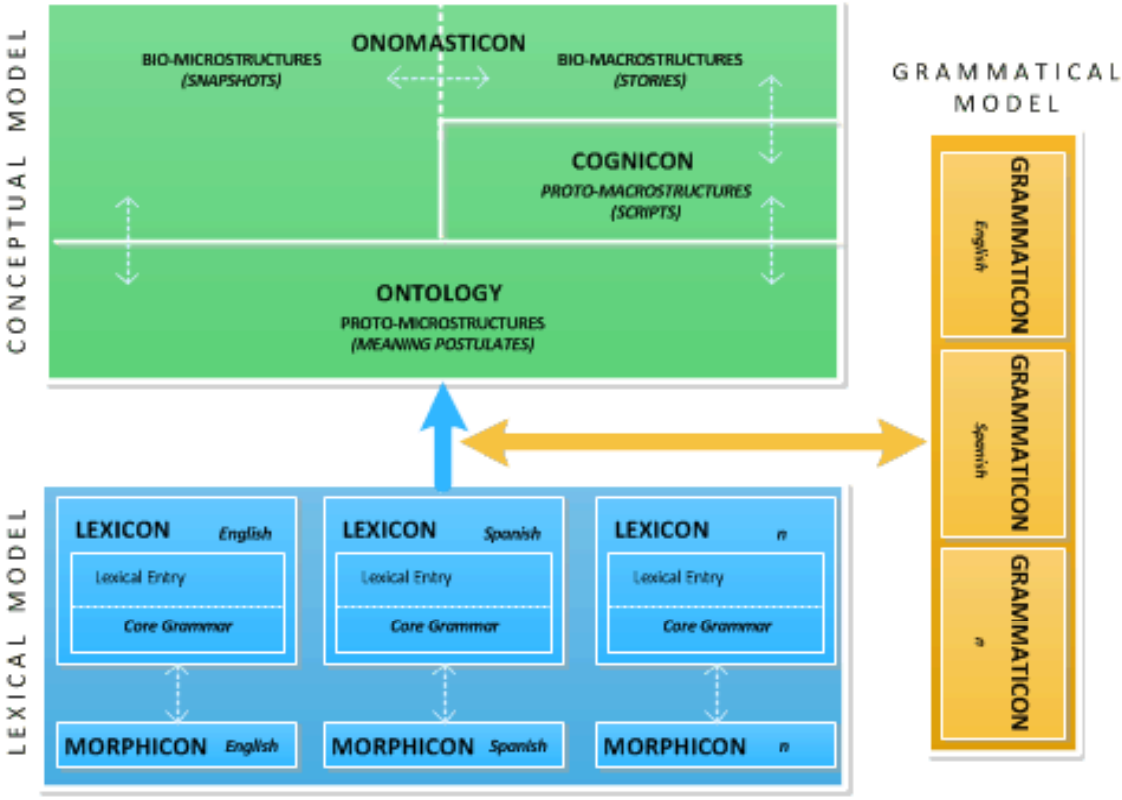

Figura 1. La suite FunGramKB (http://www.fungramkb.com)

Por tanto, podemos decir que *FunGramKB* tiene carácter conceptualista, ya que la Ontología nuclear sustenta toda la estructura de la base de conocimiento, frente a otras bases léxicas en las que el significado se expresa a través de relaciones superficiales entre unidades léxicas (por ejemplo en FrameNet o MultiWordnet).

Así pues, dado el propósito general de la Ontología nuclear ésta puede ser ampliada y enriquecida con conocimiento experto mediante enlaces con ontologías satélite (Faber, Mairal y Magaña, 2011), como se muestra en la Figura 2.

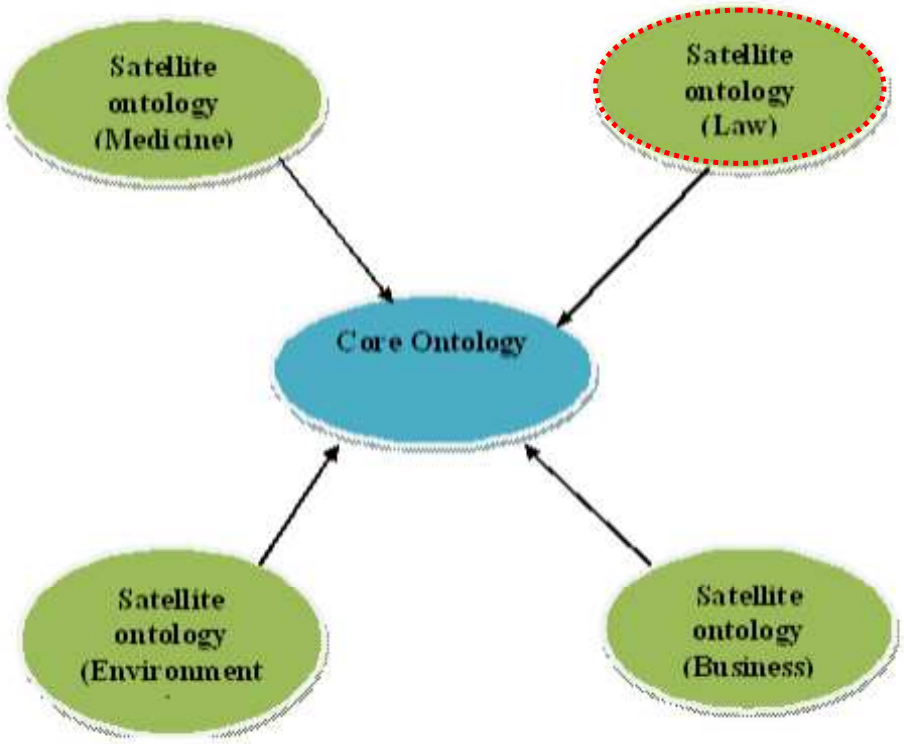

Figura 2. La ontología nuclear y las ontologías satélite (Mairal, Periñán y Samaniego, 2011: 5)

Por otro lado, la estructura multinivel de la Ontología nuclear, dividida en metaconceptos, conceptos básicos y terminales reflejan la representación conceptual de la estructura cognitiva humana.

En este estudio mostramos cómo se desarrolla la jerarquización cognitiva de la entidad +CYBERCRIME_00 en la ontología satélite del crimen organizado y el terrorismo que estará vinculada a la Ontología nuclear de *FunGramKB*.

Por ende, el desarrollo del contenido de este trabajo se estructura del siguiente modo: Primero, presentamos *FunGramKB Term Extractor* (Felices, Ureña y Alameda, 2011), herramienta para la extracción automática de términos candidatos. Seguidamente, exponemos los aspectos principales del análisis terminológico claves para la definición de términos, explicamos la metodología *COHERENT*; y finalmente se efectúa una representación de la propuesta de jerarquización de la entidad +CYBERCRIME_00.

2. FUNGRAMKB TERM EXTRACTOR

FunGramKB Term extractor es una herramienta semiautomática que asiste al terminólogo en la selección de términos ganadores. En una primera fase se obtiene de forma automática una lista de términos candidatos y se descartan los términos falsos con índice de frecuencia tf-idf menor de tres. En una fase posterior, el terminólogo realiza un primer filtrado manual de unidades léxicas (unigramas) y sintagmáticas no terminológicas (bigramas y trigramas). Posteriormente, el terminólogo realiza un segundo y último filtrado para obtener una lista de términos ganadores.

Mediante un corpus de textos, el extractor permite obtener automáticamente una lista de términos candidatos representativos de un dominio concreto que sirven al terminólogo para elaborar manualmente el filtrado de los términos y la definición de conceptos. Dichos conceptos conformarán la ontología satélite.

2.1. Conceptualización de términos ganadores

A partir de la selección de términos ganadores, el terminólogo realiza las tareas de jerarquización y conceptualización. Para ello identifica los conceptos básicos (i.e. las palabras definitorias del dominio temático) que servirán de base para la definición de los conceptos terminales, esto es, los conceptos más específicos.

Mediante el lenguaje de representación *COREL (Conceptual REpresentation Language)*[3] se modela el significado de los conceptos de la Ontología nuclear. Así, los conceptos pertenecen a tres niveles en la jerarquía conceptual de *FunGramKB*. El nivel superior lo forman 42 metaconceptos en mayúsculas precedidas por el signo "#" y representan dimensiones cognitivas. Éstas son fruto del análisis de algunas de las ontologías lingüísticas más relevantes como SUMO (Niles y Pease, 2001a, 2001b) o DOLCE (Gangemi et al., 2002; Masolo et al., 2002). La Ontología nuclear de *FunGramKB* contiene a su vez tres subontologías cuyos metaconceptos son *#ENTITY, #EVENT* y *#QUALITY* (figura 3), que permiten la organización interna de nombres, verbos y adjetivos respectivamente.

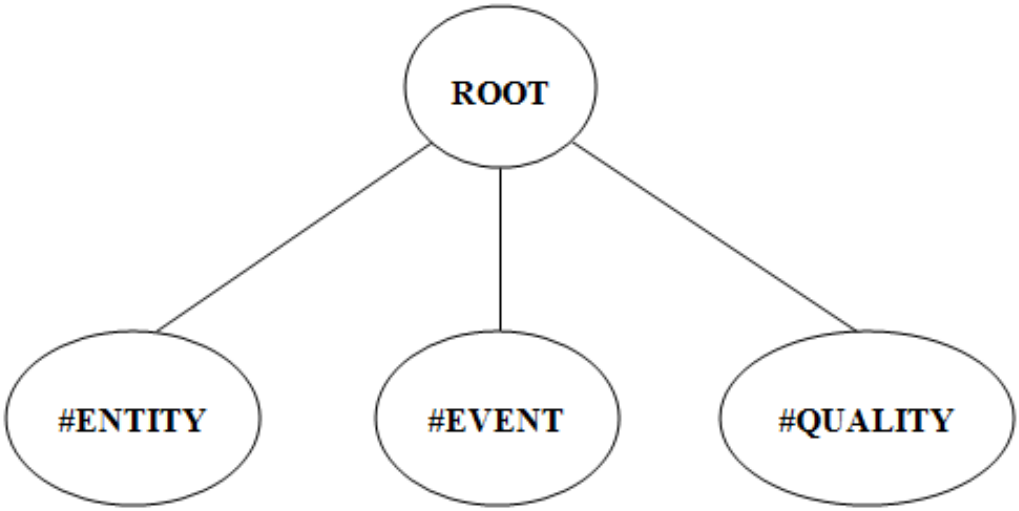

Figura 3. Tipos de conceptos

Seguidamente, en un nivel intermedio encontramos los conceptos básicos representados con el signo "+" y seguidos de un guión bajo y un índice numérico (por ejemplo, +PUNISHMENT_00, +VIOLENCE_00, etc, que podemos ver en la figura 4); y, por último están los conceptos terminales, precedidos por el signo "$" y también seguidos de un guión bajo y de un índice numérico (por ejemplo, $PENALTY_00, $RIOT_00, etc).

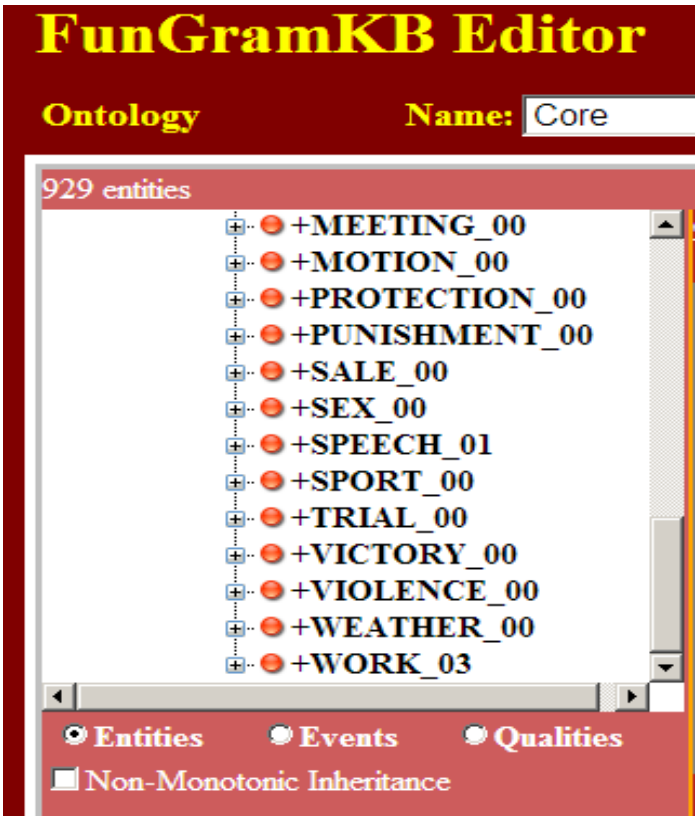

Figura 4. Representación parcial de las entidades

3. LA METODOLOGÍA COHERENT

La metodología *COHERENT* (*COnceptualization* + *HiErarchization* + *REmodelling* + *refinemeNT)* diseñada por Periñán y Mairal (2011), se utilizó para la construcción del nivel conceptual básico de la Ontología nuclear de *FunGramKB* y también sirve de base en el desarrollo de ontologías satélite (Carrión, en prensa).

En un primer paso se identificaron los conceptos básicos del *Longman Defining Vocabulary (LDV)* del *Longman Dictionary of Contemporary English* (Procter, 1978), el cual se ha probado como referencia punto de referencia en el desarrollo del vocabulario básico de un lenguaje artificial. Sin embargo, hubo que realizar una revisión profunda para realizar el mapa conceptual. Concretamente, tanto la población como la estructuración del nivel conceptual básico de la

Ontología nuclear se desarrollaron manualmente siguiendo la metodología *COHERENT* en las cuatro fases que se muestran en la Figura 5.

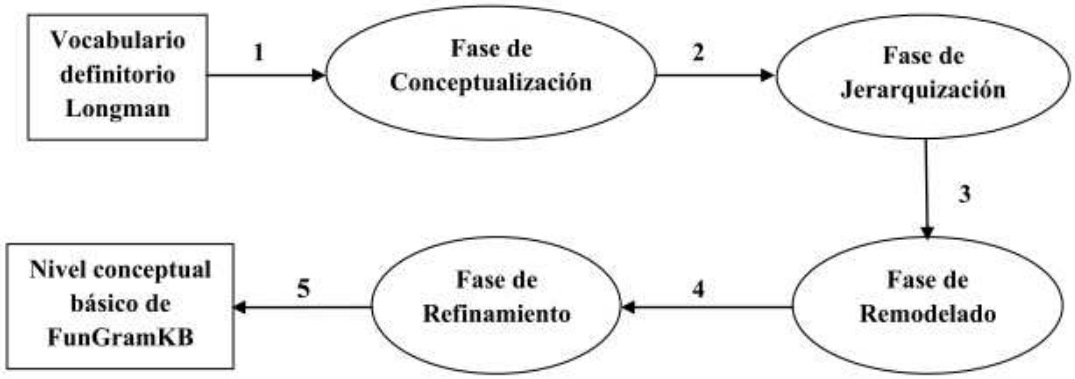

Figura 5. La metodología *COHERENT* (adaptado de Periñán y Mairal, 2011:20)
(1) Lista de unidades léxicas en inglés.
(2) Inventario de unidades conceptuales en varias lenguas.
(3) Taxonomía jerárquica de conceptos básicos incluyendo sus Marcos Temáticos (MT) y Postulados de Significado (PS).
(4) Taxonomía conceptual que incluye a los subconceptos.
(5) Nivel básico refinado de la ontología nuclear.

Como resultado de estas cuatro fases se obtuvo un catálogo de aproximadamente 1.300 conceptos básicos que han servido de base para poblar la Ontología de conceptos terminales, proceso que sigue todavía en desarrollo.

3.1. La jerarquización cognitiva

La jerarquía conceptual se establece de acuerdo a la relación de subsunción (*IS-A*) que permite la herencia múltiple no monotónica, donde los conceptos están relacionados con las unidades léxicas, pero no dependen de ellas. Así, la definición en *COREL* está formada por el GENUS+DIFFERENTIAE, donde el GENUS es compartido por sus conceptos hipónimos.

Para ilustrar esto presentamos y analizamos un par de conceptos extraídos de la jerarquía conceptual de +CYBERCRIME_00. Para ello comenzamos por definir *cybercrime* y después presentamos dos conceptos hipónimos de éste.

La definición de *cybercrime* que presentamos a continuación es fruto de la síntesis de diversas fuentes lexicográficas consultadas.

(1) *Cybercrime*: A crime in which a person uses a computer to damage or destroy equipment or to steal electronically stored data. Crime committed over the Internet.

Conceptualización en la ontología satélite:

+CYBERCRIME: A crime in which a person uses a computer to damage or destroy equipment or to steal electronically stored data. Crime committed over the Internet.

+(e1: +BE_00 (x1: +CYBERCRIME_00)Theme (x2: +CRIME_00)Referent)
+(e2: +DO_00 (x3: +HUMAN_00)Theme (x2)Referent (f1: +THE_INTERNET_00)Location)
+(e3: +USE_00 (x3)Theme (x4: +COMPUTER_00)Referent (f2)Instrument (f3: (e4: +DAMAGE_00 ^ +DESTROY_00 (x3)Theme (x5: +MATERIAL_00)Referent (f4)Instrument))Purpose)
*(e5: +STEAL_00 (x3)Theme (x6: +INFORMATION)Referent

Unidades léxicas asociadas (inglés y español): *Cybercrime, computer crime, cibercrimen, delito cibernético.*

En el proceso de jerarquización se parte de los metaconceptos más primarios para luego llegar al concepto básico hiperónimo más inmediato. Así, la propuesta de jerarquización de este término queda descrita del siguiente modo:

#ENTITY>#PHYSICAL>#PROCESS>+OCCURRENCE_00>
+CRIME_00>+**CYBERCRIME_00**

Seguidamente analizamos los conceptos hipónimos *pharming* y *phishing* como actividades prototípicas de la delincuencia informática. Para ello ofrecemos primero su definición, síntesis de las fuentes

consultadas, después su formalización en *COREL* y finalmente su propuesta de jerarquización.

(2) *Pharming:* a hacker's attack to redirect the website's traffic to a false one.

Conceptualización en la ontología satélite:

$PHARMING_00: a hacker's attack to redirect the website's traffic to a false one.

+(e1: +BE_00 (x1: $PHARMING_00)Theme (x2: +CYBERCRIME_00)Referent)
+(e2: +ATTACK_00 (x3: +HACKER_00)Theme (x4: +WEBSITE_00)Referent (f1) Instrument (f2: (e3: +COMMAND_00 (x3)Theme (x5: +TRAFFIC_00)Referent (x6: +WEBSITE_00)Goal)Purpose)

Unidades léxicas asociadas (inglés y español): *pharming*

La propuesta de jerarquización de este término es la siguiente:

#ENTITY>#PHYSICAL>#PROCESS>+OCCURRENCE_00>
+CRIME_00>+CYBERCRIME_00>**$PHARMING_00**

(3) *Phishing:* The sending of a fraudulent electronic communication to steal an identity or sell it to another party for illegal purposes.

Conceptualización en la ontología satélite:

$PHISHING_00: The sending of a fraudulent electronic communication to steal an identity or sell it to another party for illegal purposes.

+(e1: +BE_00 (x1: $PHISHING_00)Theme (x2: +CYBERCRIME_00)

+(e2: +SEND_00 (x3: +CRIMINAL_00)Agent (x4: +E_MAIL_00)Theme (x5)Referent (x6)Goal (f1: +THE_INTERNET_00)Instrument)
+(e3: n+BE_01 (x4)Theme (x7: +TRUE_00)Attribute)
(f2: (e4: +STEAL 00 (x3)Theme (x8: +IDENTITY_00)Referent)Purpose) (f3: +PLACE_00)Origin ^ (f4: (e5: +SELL_00 (x3)Agent (x8)Theme (x9)Origin (x10)Goal)Purpose)

Unidades léxicas asociadas (inglés y español): *phishing*

Por último ofrecemos la propuesta de jerarquización de este término:

#ENTITY>#PHYSICAL>#PROCESS>+OCCURRENCE_00>
+CRIME_00>+CYBERCRIME_00>**$PHISHING_00**

4. CONCLUSIONES

El ejemplo de jerarquía conceptual aquí mostrado nos ha servido para probar la premisa inicial, que el enfoque conceptualista de la base de conocimiento *FunGramKB* nos permite reutilizar la Ontología nuclear en el desarrollo de ontologías satélite, a la vez que el conocimiento experto formalizado en el lenguaje de interfaz *COREL* nos sirve para enriquecer la Ontología nuclear.

NOTAS

[1] *FunGramKB* ha sido diseñada para trabajar con siete lenguas: alemán, búlgaro, catalán, español, francés, inglés e italiano.

[2] Este trabajo forma parte del proyecto de investigación denominado "Elaboración de una ontología 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 Economía y Competitividad. Código: FI2010-15983.

[3] Véase Periñán y Mairal (2010) para una descripción detallada del lenguaje *COREL*.

REFERENCIAS BIBLIOGRÁFICAS

Carrión, M. (En prensa). "Extracción y análisis de unidades léxico-conceptuales del dominio jurídico: un acercamiento metodológico desde FunGramKB", en *Revista Electrónica de Lingüística Aplicada*.

Faber, P., R. Mairal y P. Magaña 2011. "Linking a Domain-Specific Ontology to a General Ontology". *Proceedings of the 24th International Flairs (Florida ArtificalIntelligenceResearchSociety) Conference*. AAAI Press (Association for the Advancement of Artificial Intelligence).

Felices, A., P. Ureña y A. Alameda. 2011. "FunGramKB y la adquisición terminológica". *Anglogermánica Online* 2011: 66-86.

Gangemi, Aldo et al. 2002. "Sweetening ontologies with DOLCE". En: Asunción Gómez-Pérez y Richard Benjamins (eds.) *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*: 13th International Conference, EKAW 2002, Sigüenza, 1-4.

Mairal, R. y C. Periñán. 2009. "The anatomy of the lexicon component within the framework of a conceptual knowledge base". *Revista Española de Lingüística Aplicada* 22: 217-244.

Mairal, R., C. Periñán y E. Samaniego. 2011. "Using ontologies for terminological knowledge representation: a preliminary discussion". *Technological innovation in the teaching and processing of LSPs: Proceedings of TISLID'10* (eds. N. Talaván, E. Martín Monje y F. Palazón): 267-280. UNED: Madrid.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, A. Oltramari y L. Scheider. 2002. "The WonderWeb Library of Fundational Ontologies and the DOLCE Ontology". *WonderWeb Deliverable D18*. Disponible en línea. Último acceso 30/12/2012.
[http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf].

Niles, Ian y A. Pease 2001a. Origins of the Standard Upper Merged Ontology: a proposal for the IEEE Standard Upper Ontology. En: *Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*. Seattle.

Niles, Ian y A. Pease 2001b. Towards a Standard Upper Ontology. En: *Proceedings of the Second International Conference on Formal Ontology in Information Systems*. Ogunquit.

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

Periñán, C. y F. Arcas. 2010. "The architecture of FunGramKB", en *Proceedings of the Seventh International Conference on Language Resources and Evaluation.* 17-23 mayo 2010, Valeta (Malta). European Language Resources Association (ELRA): 2667-2674.

Periñán, C. y F. Arcas. 2011. "Introducción a FunGramKB". *Anglogermánica Online* 2011: 1-15.

Periñán, C. y R. Mairal. 2010. "La gramática de COREL: un lenguaje de representación conceptual". *Onomázein* 21: 11-45.

Periñán, C. y R. Mairal. 2011. "The COHERENT methodology in FunGramKB". *Onomázein* 24: 13-33.

Procter, P. (ed.). 1978. *Longman Dictionary of Contemporary English.* Harlow (Essex): Longman.