Languages for Special Purposes in a Multilingual, Transcultural World

Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria

http://lsp2013.univie.ac.at/proceedings

El conocimiento cultural en dominios especializados: Un acercamiento desde la base de conocimiento FunGramKB

Pedro Ureña Gómez-Moreno

Cite as: Ureña Gómez-Moreno, P. (2014). El conocimiento cultural en dominios especializados: Un acercamiento desde la base de conocimiento FunGramKB. In G. Budin & V. Lušicky (eds.), *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria*. Vienna: University of Vienna, 509-514.

Publication date: July 2014

ISBN: **978-3-200-03674-1**

El conocimiento cultural en dominios especializados: Un acercamiento desde la base de conocimiento FunGramKB

Pedro Ureña Gómez-Moreno
Department of English and German Philologies, Faculty of Arts and Philosophy, University of Granada
Spain

Correspondence to: pedrou@ugr.es

Abstract. FunGramKB is a knowledge base made up of different modules for the comprehensive processing of language. The main component is the conceptual, containing both common-sense knowledge (Ontology), procedural knowledge (Cognicon) as well as knowledge about named entities representing people, places or organisations (Onomasticon). This paper draws on previous studies within FunGramKB dealing with the Onomastical component and reviews this module from the perspective of specialised discourse.

Keywords. FunGramKB, cultural knowledge, ontologies.

1. FunGramKB

FunGramKB, tal y como se ha presentado en estudios anteriores (Periñán-Pascual & Arcas-Túnez, 2004, 2005, 2007a, 2007b), se define como una base de conocimiento para el Procesamiento del Lenguaje Natural (PLN), que, por un lado, es multipropósito, ya que puede utilizarse en el desarrollo de un gran número de distintas aplicaciones relacionadas con la computación del lenguaje, y que, por otro lado, es multilenguaje, en tanto que puede trabajar con distintas lenguas, independientemente de sus características morfológicas o gramaticales (ver www.fungramkb.com). Tratar de señalar todas las ventajas de esta base de conocimiento requeriría detenerse en los detalles de su gestación y fundamentación teórica, por lo que solo nos referiremos aquí a dos características fundamentales que hacen de FunGramKB un instrumento especialmente versátil. De una parte, FunGramKB está construida sobre una arquitectura robusta que se divide en varios módulos interconectados para el procesamiento lingüístico a todos los niveles (morfológico, gramatical, construccional, conceptual, etc.). Esta arquitectura está concebida como un todo estructurado, de tal forma que estos niveles funcionan de forma secuenciada. De otra parte, y quizá lo más relevante para el propósito de este artículo, FunGramKB está diseñada para integrar de forma eficaz el conocimiento del denominado "sentido común", con el conocimiento especializado, es decir, con conceptos que son propios de áreas como la Medicina, el Derecho o la Ingeniería. A continuación, se muestra de manera gráfica la arquitectura de FunGramKB:

Como se observa en la Fig. 1, en esta arquitectura existen tres niveles o "modelos" claramente diferenciados: el conceptual (marcado en verde), el léxico (en azul) y el gramatical (en amarillo). Las flechas de trazo grueso así como las discontinuas hacen referencia a la naturaleza interconectada de cada uno de los modelos y de sus subniveles de descripción, respectivamente. Por razones de espacio, este artículo no abordará una revisión de cada uno de los componentes de FunGramKB, sino que se refiere el lector a Periñán-Pascual (2007).

Este artículo explora aspectos teóricos y prácticos de la construcción del "Onomasticón" (en inglés "Onomasticon"), que forma parte del modelo conceptual. En concreto, plantea varias propuestas para el desarrollo de entidades relacionadas con ámbitos especializados de conocimiento y con el conocimiento enciclopédico. Para ello, el artículo parte del estudio preliminar de Periñán-Pascual & Carrión-Varela (2011), que constituye el primer análisis en profundidad del componente onomástico de FunGramKB. Tal y como se verá, es posible orientar el Onomasticón para su aplicación en tareas de procesamiento y razonamiento artificial.

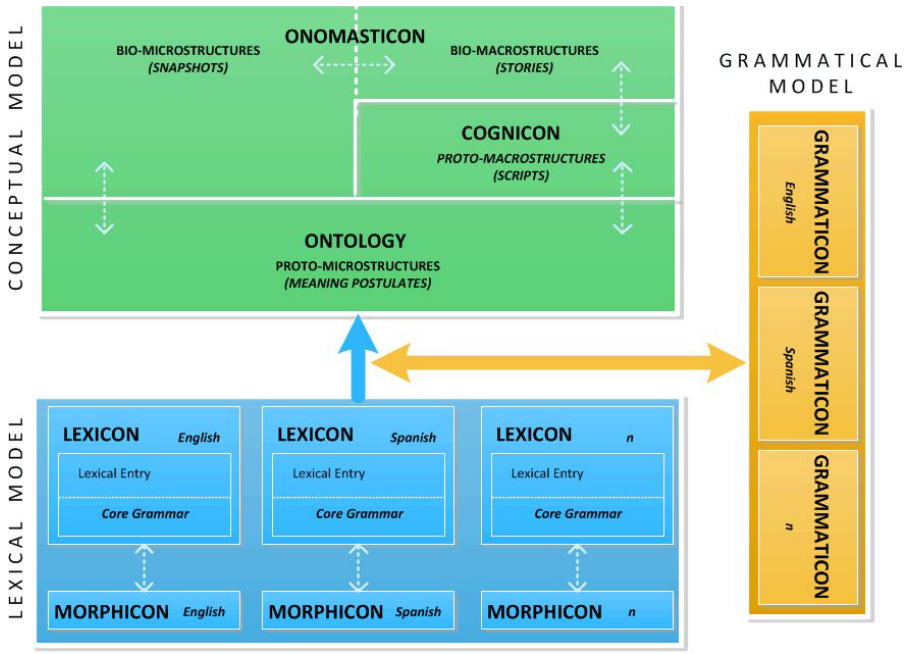

Figure 1: La arquitectura de la base de conocimiento FunGramKB

El resto de este artículo se estructura de la siguiente forma. El segundo apartado ofrece una revisión de los aspectos más relevantes del Onomasticón. El tercer apartado aborda cuestiones metodológicas relacionadas con la conceptualización en el Onomasticón e ilustra el proceso mediante algunos ejemplos pertenecientes a los dominios especializados del terrorismo y el crimen organizado. Finalmente, el cuarto apartado plantea pautas generales para el aprovechamiento y la aplicación del conocimiento onomástico en tareas de razonamiento y descubrimiento de información.

2. El Onomasticón

El Onomasticón se define como un módulo conceptual para el registro de unidades de tipo enciclopédico, es decir, unidades conceptuales que representan entidades y eventos únicos (Periñán-Pascual & Carrión-Varela 2011). Por ejemplo, el Onomasticón contiene información sobre personas como *Carlos V* o *Nelson Mandela*, o lugares como *Viena* o *La Alhambra*. Al igual que la ontología (véase Fig. 1), los conceptos del Onomasticón están definidos mediante lenguaje COREL (COnceptual REpresentation Language) (Periñán-Pascual & Mairal-Usón 2010), que permite expresar información conceptual de una manera flexible. Este lenguaje además posee la ventaja de ser procesable computacionalmente a la vez que fácilmente legible para los humanos. El Onomasticón (véase Fig. 1) contiene dos tipos de unidades conceptuales: las bio-microestructuras y las bio-macroestructuras. Las primeras capturan conocimiento cultural síncrono, mientras que las segundas representan historias de forma diacrónica (Periñán-Pascual 2012). En este apartado se aborda el estudio de dos bio-microestructuras de tipo onomástico.

Aprehender total o parcialmente el conocimiento del mundo así como describir las entidades conocidas son tareas complejas que requieren tiempo. De ahí que la población del Onomasticón se lleve a cabo primordialmente de forma semiautomática mediante la reutilización de conocimiento proveniente de otras fuentes de información digitales abiertas como, por ejemplo, Wikipedia. El Onomasticón de FunGramKB puede importar el conocimiento compilado en estos repositorios digitales para conectarlo con el resto de modelos y aumentar así la capacidad de procesamiento y razonamiento de la base de conocimiento. En el caso de Wikipedia, por ejemplo, este proceso de importación puede realizarse desde los "cuadros de información" (en inglés *infoboxes*) al Onomasticón utilizando para ello la mediación de la taxonomía de DBpedia. DBpedia es el producto de la colaboración entre grupos de usuarios y se ofrece como un repositorio ontológico

formado a partir de la información estructurada de Wikipedia. El proceso de importación de DBpedia al Onomasticón se divide en tres fases (Periñán-Pascual & Carrión-Varela 2011). En primer lugar, se construyen las plantillas COREL que servirán como receptáculo conceptual de los datos provenientes de la base de datos enciclopédica. Esta fase se lleva a cabo a través de una interfaz específica integrada en la plataforma de edición de FunGramKB. En segundo lugar, se realiza el volcado estructurado de la información desde DBpedia al Onomasticón. Finalmente, en la última fase se realizan tareas de mantenimiento de las plantillas así como de actualización de los datos enciclopédicos, ya que la información importada no es estática, sino que puede sufrir variaciones con el paso del tiempo. Esta metodología resulta de gran utilidad para la adquisición masiva de nueva información. Sin embargo, no incluye por el momento la adquisición de contenidos no estructurados, es decir, datos que no están organizados bajo un patrón recurrente como el que ofrecen los cuadros de información, sino que se encuentran desarrollados sin forma predefinida en el cuerpo principal de la entrada enciclopédica. En la siguiente sección, se proponen algunos ejemplos de conceptualización de unidades onomásticas a partir de la información no estructurada de Wikipedia.

3. Instancias especializadas en el Onomasticón

FunGramKB es una base de conocimiento en evolución que se somete a continuas tareas de mejora, expansión y filtrado de información en todos sus niveles. Uno de los principales avances de la base de conocimiento ha sido la reestructuración de su diseño con el objetivo de albergar las denominadas "Ontologías Satélite", es decir, ontologías de contenido especializado construidas sobre la base de terminología propia de campos científicos o técnicos (para un estudio en profundidad la construcción de ontologías terminológicas, se refiere el lector a Felices & Ureña 2011, Ureña et al. 2011, Carrión-Delgado 2012, Felices & Ureña 2012). Estas ontologías especializadas están conectadas a la Ontología Nuclear y ésta a su vez conecta las Ontologías Satélite entre sí.

Separar entre Onomasticón especializado y no especializado en los mismos términos que se ha hecho entre la Ontología Nuclear y las Ontologías Satélites no es una tarea sencilla, ya que el Onomasticón alberga conocimiento no prototípico o no estereotípico (Periñán-Pascual 2012). De hecho, la arquitectura de FunGramKB no contempla especificación alguna en este sentido (véase Fig. 1). Para nuestro propósito, no obstante, establecemos esta distinción a efectos puramente metodológicos y definimos conceptos culturales especializados como unidades relevantes en tareas de procesamiento o de recuperación de información en un dominio de conocimiento técnico. Esta sección se centra en algunos ejemplos de conceptualización de unidades especializadas del dominio criminal, en concreto en relación con la lucha contra el crimen organizado y el terrorismo, y ofrece una reflexión sobre cómo esta conceptualización puede ser útil en tareas de procesamiento.

Una primera distinción en la conceptualización del dominio del crimen internacional consiste en la división entre agentes criminales, esto es, entidades que realizan actos de delincuencia organizada o terrorista, y entidades de defensa, que ejercen labores de prevención o captura de los agentes criminales. Otras instancias necesarias para el modelado del dominio incluyen los materiales, las técnicas y las armas que ambos tipos de agentes utilizan para la prevención o, por contrario, comisión de los delitos. A continuación se muestra un ejemplo la conceptualización de tipo cultural del organismo EUROPOL, entidad europea para la prevención de los delitos antes mencionados. La información que aparece abajo en cursiva está extraída de Wikipedia (versión en lengua inglesa) y bajo ésta se muestra la correspondiente representación COREL:

(1) EUROPOL

- *Europol's aim is to improve the effectiveness and co-operation between the competent authorities of the member states primarily by sharing and pooling intelligence to prevent and combat serious international organized crime.*

*(e1: +BE_00 (x1: %EUROPOL_00)Theme (x2: +ORGANIZATION_00)Referent)

*(e2: +HELP_00 (x1)Theme (x3: %EUROPEAN_UNION_00)Referent (f1: (e3: n +EXIST_00 (x4: +ORGANIZED_CRIME_00)Theme))Purpose)

- Dismantling of a credit card fraud network.

*((e4: past +FINISH_00 (x1)Theme (x5: +ORGANIZATION_00)Referent) (e5: +DO_00 (x5) Theme (x6: +FRAUD_00)Referent (x7: $CREDIT_CARD_00)Instrument))

- Seizing of a 30 kilograms (70 lb) cocaine load from Colombia

*(e6: past +SEIZE_00 (x1)Theme (x8: $COCAINE_00)Referent (f2: 30 $KILOGRAM_00) Quantity (x9)Origin (x10)Goal (x11: %COLOMBIA_00)Location)

- Disruption of an illegal immigration network in France

*(e7: past +FINISH_00 (x1)Theme (x12: $IMMIGRATION _00)Referent (x13: %FRANCE_00) Location (f3: (e8: n +BE_00 (x12)Theme (x14: $LEGAL_N_00)Attribute))Condition)

La primera de las proposiciones mostradas arriba pertenece al epígrafe "Funciones" (*Functions*) que se encuentra bajo la entrada principal de EUROPOL, mientras el resto de predicaciones pertenecen a la sección dedicada a las "Operaciones" (*Operations*) de este organismo. Tal y como se ha sugerido anteriormente, la información en el cuerpo de los artículos de Wikipedia no es de tipo estructurado en el mismo sentido que los cuadros de información y, por tanto, su compilación en FunGramKB es a día de hoy manual. Para incorporar este conocimiento sin estructura explícita en FunGramKB es necesario realizar dos pasos principalmente. En primer lugar, seleccionar exclusivamente la información necesaria y, en segundo lugar, traducir la información seleccionada a lenguaje COREL. Esta traducción y representación formal en COREL habrá de simplificar las estructuras más complejas del lenguaje natural (véase, por ejemplo, la reducción entre la primera predicación en el ejemplo (1) y su correspondiente traducción en COREL).

Un segundo ejemplo de modelado de un concepto onomástico del crimen organizado atañe a las personas que han ejercido la delincuencia. En el caso del conocido como Al Capone encontramos que existe una gran número de datos relativos a este personaje en Wikipedia. Como ocurre en el caso anterior, la mayoría de la información a esta entidad no se encuentra en una cuadro de contenido, sino desarrollado en el cuerpo del artículo. A continuación se analiza la versión española de la entrada enciclopédica de Al Capone y, como se aprecia, el lenguaje de conceptualización es el mismo que en el caso anterior. Este caso se propone como ejemplo de perfil criminal y podría servir como modelo de conceptualización para otros perfiles criminales actuales:

(2) ALCAPONE

- Alphonse Gabriel Capone […] más conocido como Al Capone o Al Scarface Capone […], apodo que recibió debido a la cicatriz que tenía en su cara, provocada por un corte de navaja.

*(e1: past +BE_00 (x1: %AL_CAPONE)Theme (x2: %SCARFACE_00)Referent (f1: (e2: +HAVE_00 (x1)Theme (x3: $SCAR_00)Referent (x4: +FACE_00)Location))Reason)

- […] Capone siguió enriqueciéndose gracias al tráfico ilegal de bebidas alcohólicas […], y a través de su vasta red clandestina de salas de juego.

*(e3: past +OBTAIN_00 (x1)Theme (x5: +MONEY_00)Referent (f2: (e4: +SELL_00 (x1) Theme (x6: +BEVERAGE_00)Referent (x7: $LEGAL_N_00)Attribute))Means

*(e5: +CREATE_00 (x1)Theme (x8: i +ROOM_00)Referent (f3: (e6: +PLAY_00 (x9)Theme (x10)Referent (x11: $LEGAL_N_00)Attribute))Purpose)

- Aunque probablemente nunca fue iniciado en la Cosa Nostra, rápidamente se asoció

con la Mafia y se adueñó del hampa de Chicago

*(e7: past n +HAVE_00 (x1)Theme (x12: %COSA_NOSTRA_00)Referent)

*(e8: past +JOIN_00 (x1)Theme (x13: %MAFIA_00)Referent)

*(e9: past +BE_00 (x1)Theme (x14: +LEADER_00)Referent (x15: $HAMPA_00)Beneficiary)

Uno de los objetivos a medio plazo en el desarrollo de FunGramKB es la asimilación semiautomática de conocimiento no estructurado disponible en la red o soporte informático. No obstante, la base de conocimiento que proponemos ya ha dado un gran paso en esta dirección al plantear una primera implementación de ARTEMIS (Automatically Representing TExt Meaning via an Interlingua-based System), que servirá como interfaz de computación textual de intermediación entre el lenguaje natural y la propia base de conocimiento. Una de las muchas vías de investigación que abre ARTEMIS es precisamente la de mejorar la forma en la que FunGramKB adquiere conocimiento. Por otro lado, este sistema interlingüístico también contribuirá a mejorar las funciones de razonamiento, tanto cualitativa como cuantitativamente.

4. Conocimiento cultural y razonamiento artificial

Las labores de adquisición y modelado conceptual que se está realizando en el marco de FunGramKB están encaminadas a la computación del lenguaje y al razonamiento artificial. Los procesos de razonamiento en esta base de conocimiento están divididos en dos: por un lado, el "Microknowing", o proceso de nivel bajo para la herencia e inferencia conceptuales, y, por otro lado, el "Macroknowing", encargado de establecer conexiones e inferencias entre el conocimiento ontológico, onomástico y procedimental (Periñán-Pascual & Arcas-Túnez 2005; Periñán-Pascual & Mairal-Usón 2009). A nivel más general, la introducción de mecanismos de razonamiento conceptual como el que está desarrollando FunGramKB va a permitir que los organismos y las instituciones que desempeñan su labor en distintas áreas de trabajo puedan utilizar la base de conocimiento para acceder a información relevante de forma más rápida e inteligente.

El papel que juega el conocimiento cultural del mundo representado por las entidades y por las historias contenidas en el Onomasticón resulta fundamental para establecer relaciones entre entidades o eventos aparentemente no relacionados. En el caso de los dominios del crimen organizado y terrorismo, por ejemplo, la descripción de las células y agentes colectivos o individuales que llevan a cabo actos delictivos, puede contribuir en el descubrimiento y predicción de nuevos perfiles de riesgo relacionados con personas o grupos que muestren pautas similares de comportamiento delictivo. De esta forma, eventos como "participación", "pertenencia", "colaboración" o "compra-venta", etc. referidos a una entidad cualquiera pueden poner a ésta en relación con otra entidad más conocida y, lo que es aún más importante, esta relación podrá sugerir acciones o recomendar toma de decisiones. El motor de razonamiento de la base de conocimiento podrá emplearse asimismo para encontrar relaciones entre los lugares de comisión de delitos y personas bajo investigación.

5. Conclusiones

Este artículo ha ofrecido una revisión general de la base de conocimiento FunGramKB y ha tratado de discutir dos aspectos fundamentalmente. En primer lugar, se ha propuesto avanzar en el desarrollo del componente de FunGramKB denominado "Onomasticón" –en lo que se refiere en particular al modelado conceptual de entidades y eventos relativos al terrorismo y el crimen organizado– tomando como punto de partida la importación de datos desde otras fuentes de conocimiento, así como la creación *ad hoc* de información. Tal y como se ha mencionado, gran parte de la información cultural (y no cultural) de la que disponemos en los distintos repositorios de datos en formato digital aparece expresada de forma no estructurada, de ahí que actualmente la población del Onomasticón se lleve a cabo principalmente de forma automática a partir de la

importación desde bases de datos de contenido estructurado. Sin embargo, este proceso podrá automatizarse parcialmente en fases posteriores de desarrollo para agilizar la incorporación de nuevas unidades conceptuales a partir de datos no estructurados. En segundo lugar, este artículo ha propuesto de forma preliminar algunas vías de aplicación del conocimiento onomástico más especializado en labores de razonamiento y descubrimiento de información. Tantos los presupuestos teóricos como metodológicos para crear unidades conceptuales en el Onomasticón se proponen *a priori* para cualquier campo especializado de conocimiento.

6. 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.

7. Referencias

Carrión-Delgado, M. de Gracia (2012). Extracción Y Análisis De Unidades Léxico-Conceptuales Del Dominio Jurídico: Un Acercamiento Metodológico Desde FunGramKB. *RaeL 11*, 25-39.

DBpedia: http://dbpedia.org/.

Felices-Lago, Ángel, & Ureña Gómez-Moreno, Pedro (2011). FunGramKB Y La Adquisición Terminológica. *Anglogermánica Online: Electronic Journal of English and German Philology 1-2011*, 66-86.

Felices-Lago, Ángel, & Ureña Gómez-Moreno, Pedro (2012). Fundamentos Metodológicos De La Creación Subontológica En FunGramKB. *Onomázein 26*, 49-67.

FunGramKB: A lexico-conceptual knowledge base for NLP. (http://www.fungramkb.com/).

Periñán-Pascual, Carlos (2012). The Situated Common-sense Knowledge In FunGramKB. *Review of Cognitive Linguistics 10* (1), 184-214.

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

Periñán-Pascual, Carlos, & Arcas-Túnez, Francisco (2007a). Cognitive Modules Of An NLP Knowledge Base For Language Understanding. *Procesamiento del Lenguaje Natural 39*, 197-204.

Periñán-Pascual, Carlos, & Arcas-Túnez, Francisco (2007b). Deep semantics in an NLP knowledge base. 1*2th Conference of the Spanish Association for Artificial Intelligence*, Universidad de Salamanca, 279-288.

Periñán-Pascual, Carlos, & Arcas-Túnez, Francisco (2010). Ontological Commitments In FunGramKB. *Procesamiento del Lenguaje Natural 44*, 27-34.

Periñán-Pascual, Carlos, & Carrión-Varela, María de los Llanos (2011). FunGramKB Y El Conocimiento Cultural. *Anglogermánica 1-2011*, 87-105.

Periñán-Pascual, Carlos, & Mairal-Usón, Ricardo (2009). Bringing Role and Reference Grammar To Natural Language Understanding. *Procesamiento del Lenguaje Natural 43*, 265-273.

Periñán-Pascual, Carlos, & Mairal-Usón, Ricardo (2010). La Gramática De COREL: Un Lenguaje De Representación Conceptual. *Onomázein 21*, 11-45.

Ureña Gómez-Moreno, Pedro, Alameda-Hernández, Ángela, & Felices-Lago, Ángel (2011). Towards a specialised corpus of organized crime and terrorism. In María Luisa Carrió et al. (eds.) *La investigación y la enseñanza aplicadas a las lenguas de especialidad y a la tecnología*. Universitat Politècnica de Valencia, Valencia, 301-306.

Wikipedia: http://es.wikipedia.org/. Artículos mencionados: EUROPOL (http://es.wikipedia.org/wiki/Europol) (Acceso Agosto 2013) y Al Capone (http://es.wikipedia.org/wiki/Al_Capone) (Acceso Agosto 2013).