

RESOLUCIÓN DE ANÁFORAS QUE REQUIEREN CONOCIMIENTO CULTURAL CON LA HERRAMIENTA FUNGRAMKB*

María de los Llanos Carrión Varela

Universidad Nacional de Educación a Distancia (Madrid)

Abstract: La integración de diversos tipos de conocimiento lingüístico en sistemas de comprensión o procesamiento del lenguaje natural (PLN) constituye una práctica común. Sin embargo, habitualmente se ha obviado la inclusión de conocimiento cultural, ya sea por motivos técnicos o teóricos. No obstante, un sistema del PLN enriquecido con información cultural constituye una herramienta más robusta y cohesionada, para llevar a cabo más óptimamente la resolución de problemas vinculados a la comprensión del lenguaje natural como, por ejemplo, la ambigüedad existente en fenómenos lingüísticos como la anáfora, referencia y correferencia o metáfora y metonimia, entre otros. El objetivo de este artículo es describir cómo la base de conocimiento FunGramKB integra el conocimiento cultural a través de sus módulos conceptuales y, en especial, cómo la información del módulo denominado *Onomasticón* puede contribuir a maximizar la informatividad del sistema completo, para resolver problemas de ambigüedad en un determinado fenómeno lingüístico: la anáfora.

Palabras clave: PLN, FunGramKB, resolución de anáfora, bases de conocimiento, ingeniería del conocimiento.

Abstract: While integrating linguistic knowledge of any kind is becoming an almost implicit practice in natural language understanding systems, the inclusion of cultural or world knowledge in these tools might have been neglected sometimes. However, a NLP system or knowledge base enriched with cultural information is a more robust, better cohesioned instrument for natural language understanding processes. The integration of this type of knowledge in NLP systems may be proven to contribute to solving some phenomena that occur in natural language, such as anaphor, metaphor and metonymy, ambiguity or co-reference, amongst others. The objective of this article is to describe the way FunGramKB (a knowledge base) integrates cultural knowledge in its conceptual modules and, in particular, how the information contained in the *Onomasticon* module of FunGramKB can contribute to maximising the informativeness and completeness of the whole system, thus resolving ambiguity problems in a determined linguistic phenomenon: anaphora.

Keywords: NLP, FunGramKB, anaphora resolution, knowledge bases, knowledge engineering.

1. INTRODUCCIÓN

1.1. La ambigüedad en el lenguaje

Un hecho importante tras el cual subyacen muchos de los problemas y cuestiones por resolver en procesamiento del lenguaje natural (en adelante, PLN) es que, para las computadoras, resulta una ardua tarea el poder entender el lenguaje natural. Esta obvia característica, derivada de la ausencia de sentido común por parte de las máquinas, se encuentra principalmente motivada o agravada por una característica innata del lenguaje natural: la ambigüedad. Esta particularidad podría quizás ser consecuencia de la evolución natural favorecida por el principio de economía en la lengua, mediante el cual el ser humano es capaz de transmitir el máximo de información posible utilizando el mínimo de signos lingüísticos. Esta ley podría impulsar la existencia de ambigüedad lingüística de varios tipos: ambigüedad semántica (de una misma lexicalización), ambigüedad sintáctica (provocada por elipsis y fenómenos análogos), o incluso ambigüedades más allá del mero texto y que aparecen a nivel pragmático (por ejemplo, utilizar una pregunta que se debe interpretar como una petición o incluso una orden).

* Este trabajo forma parte del proyecto de investigación FFI2011-29798-C02-01, financiado por el Ministerio de Ciencia e Innovación

Por tanto, la resolución de las diversas manifestaciones de la ambigüedad en el lenguaje natural constituye uno de los sempiternos campos de trabajo dentro del PLN, para lo cual intentaremos mostrar la aportación resolutoria que la base de conocimiento FunGramKB puede ofrecer. En concreto, en el presente artículo se tratará de ilustrar cómo la inclusión de conocimiento cultural en una base de conocimiento de semántica profunda como FunGramKB puede ayudar a resolver algunos de los más frecuentes problemas de comprensión en el PLN y, específicamente, uno de los fenómenos lingüísticos más comunes que guardan relación con la ambigüedad: la anáfora. Para ello, primeramente se realizará una presentación de la herramienta en la que se desarrolla el actual trabajo: la base de conocimiento FunGramKB, con un detalle más extenso del módulo dentro de la misma denominado *Onomasticón*. Una vez definida la herramienta FunGramKB y su *Onomasticón*, en los diversos epígrafes de la sección 2 se procederá a efectuar una contextualización de la problemática, mediante la descripción de diferentes fenómenos anafóricos que son de interés para el PLN, así como la muestra de varios ejemplos citados en la literatura a los cuales FunGramKB es capaz de aportar una nueva solución. Finalmente, la sección 3 se compone de las conclusiones y futuros retos de investigación que se plantean a continuación del trabajo mostrado en el presente artículo.

1.2. Qué es FunGramKB

La herramienta FunGramKB¹ (Periñán Pascual y Arcas Túnez 2004, 2007, 2008, 2010a, 2010b; Periñán Pascual y Mairal Usón, 2009, 2010; Van Valin y Mairal Usón, en prensa; Mairal Usón, 2012; Mairal Usón *et al.* 2012; Periñán Pascual y Mairal Usón, 2012) es una base de conocimiento léxico-conceptual multipropósito, creada para ser implementada en diversas aplicaciones de PLN. La base FunGramKB es multifuncional, ya que ha sido diseñada para ser utilizada en múltiples tareas de PLN, así como multilingüe, puesto que está soportada en varios idiomas².

FunGramKB está estructurada en tres grandes niveles de información, que a su vez se subdividen en varios módulos independientes pero interrelacionados (Periñán Pascual y Arcas Túnez, 2006; Periñán Pascual y Arcas Túnez, 2010b). Estos tres grandes niveles son el *nivel léxico* (conocimiento lingüístico), el *nivel gramatical* (conocimiento acerca de esquemas construccionales) y el *nivel conceptual* (conocimiento no lingüístico). La integración de módulos y tareas permite a FunGramKB ser una herramienta versátil y completa, al combinar diversos bloques que cubren tanto el aspecto lingüístico del lenguaje natural (módulos Léxico y Gramatical, adaptados a cada una de las lenguas naturales soportadas) como el aspecto cognitivo (módulos Conceptuales), que son los módulos compartidos por todas las lenguas naturales y que actúan como eje pivotal de la herramienta. La figura que se incluye a continuación refleja la arquitectura de FunGramKB:

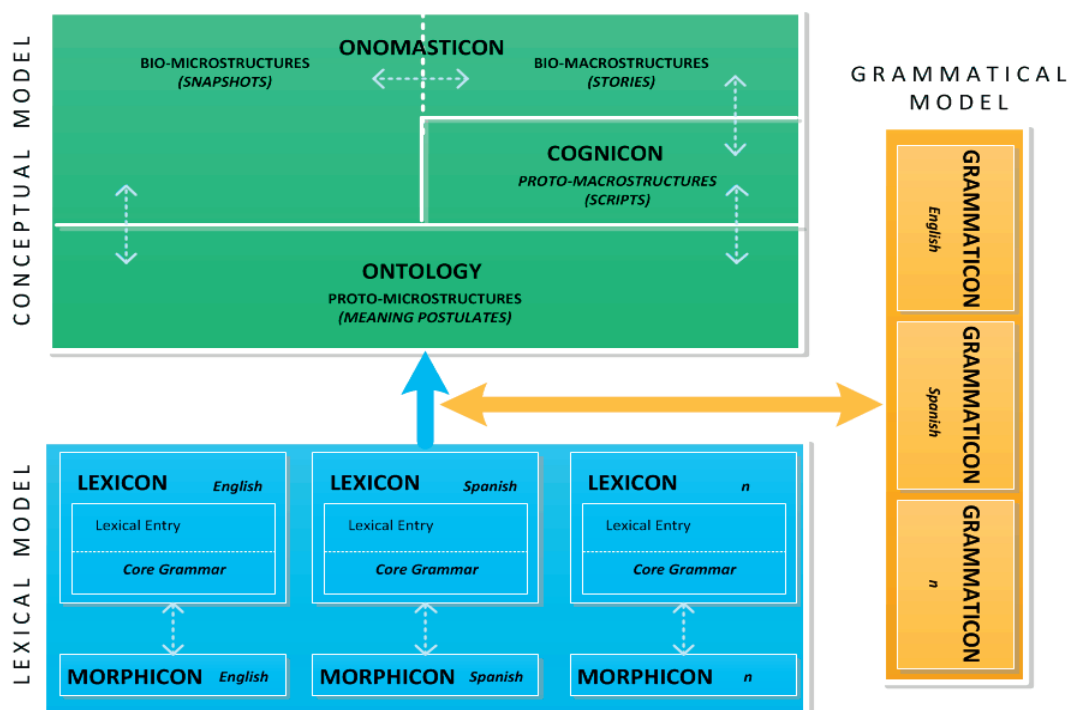


Figura 1. La arquitectura de FunGramKB³

¹ <http://www.fungramkb.com/>

² Alemán, búlgaro, catalán, español, francés, inglés e italiano.

³ Figura obtenida de <http://www.fungramkb.com/> [fecha de consulta: 14/10/2013]

Si se presta atención a los diferentes módulos lingüísticos, comenzando por el *nivel léxico*, se puede explicar sus componentes de la siguiente forma: el *Lexicón* es el módulo que contiene la información morfosintáctica, pragmática y colocacional de las unidades léxicas de cada lengua. Se trata, por tanto, de información que ayuda a utilizar dichas unidades de manera correcta formalmente en el discurso. El *Morficón* de cada lengua contiene aquellas reglas que afectan a la morfología flexiva, como por ejemplo las inflexiones verbales o conjugaciones, así como las concordancias de género y número. El siguiente módulo lingüístico, el *nivel gramatical*, contiene el denominado *Construcción*, que almacena los *Gramaticones* de las diferentes lenguas naturales, los cuales constituyen esquemas de construcciones que permiten construir la interfaz entre sintaxis y semántica, denominada enlace (*linking algorithm*), para representar un texto de entrada en lenguaje natural mediante una estructura lógica. Asimismo, en FunGramKB esta estructura lógica se ve realizada mediante un nuevo formalismo denominado “estructura lógica conceptual” (CLS por sus siglas en inglés, *Conceptual Logical Structure*), que maximiza la carga informativa y reduce la redundancia (Periñán Pascual y Arcas Túnez, 2010b).

El otro gran módulo que compone FunGramKB es el *nivel conceptual*. Este nivel almacena aquellas representaciones prototípicas de la realidad que el individuo recrea en su mente, no las palabras con las que se describe de manera lexicalizada dicha realidad. De este modo, mientras que el módulo de nivel léxico es particular para cada lengua natural (ha de crearse, pues, un *Lexicón* para el inglés, otro para el español, y así sucesivamente para cada lengua con cada uno de los tres componentes del nivel léxico), el nivel conceptual es común a todos los lenguajes naturales, puesto que no está basado en conocimiento sobre las palabras sino en conocimiento sobre el mundo (Periñán Pascual y Arcas Túnez, 2006). Es por esto que se puede afirmar que el nivel conceptual de FunGramKB es universal, en tanto en cuanto los conceptos que recopila y categoriza son comunes al mundo (Periñán Pascual y Arcas Túnez, 2007:199): “*FunGramKB ontology takes the form of a universal concept taxonomy, where ‘universal’ means that every concept we can imagine has an appropriate place in this ontology.*” Se trata de la lexicalización de estos conceptos, la manera de organizarlos y de aglutinarlos en expresiones lingüísticas en las diferentes lenguas naturales, lo que difiere entre ellas. No obstante, el hecho de que el módulo conceptual de FunGramKB sea denominado *universal* no lo exime de ser lingüísticamente motivado, pero no lingüísticamente dependiente. Esto significa que cada uno de los conceptos introducidos en el módulo conceptual (particularmente en la Ontología) tiene, necesariamente, al menos una unidad léxica cuyo significado no coincide con ninguno de los otros conceptos ya presentes en la base de conocimiento. Asimismo, este proceso asegura que cada nueva unidad léxica que pueda surgir en el futuro, como consecuencia de la introducción de nuevas lenguas naturales en FunGramKB, tendría cabida dentro de ella tras un proceso de negociación. La posibilidad de introducción de estos nuevos conceptos, lexicalizados de una manera particular en una lengua natural, además, contribuye a eliminar una posible carga subjetiva provocada por las propias lenguas maternas y contextos culturales de los ingenieros del conocimiento que realizan la tarea de poblar la base (Periñán Pascual y Arcas Túnez, 2010b).

El nivel conceptual de FunGramKB se compone de tres módulos: la *Ontología*, que es una estructura jerárquica de conceptos usados por las personas para describir cualquier situación cotidiana; el *Cognición*, que almacena conocimiento procedimental en forma de secuencias temporales o guiones, y que está basado en el modelo temporal de Allen (Allen, 1983; Allen y Ferguson, 1994); y el *Onomasticón*, que almacena conocimiento episódico y cultural acerca de entidades (p. ej., personas, ciudades, lugares, acontecimientos, etc.) en forma de bio-estructuras.

A continuación, describiremos brevemente los tres módulos del nivel conceptual de FunGramKB, puesto que éste será el módulo en el que se centrarán las tareas que mostramos en el presente trabajo, más concretamente en el Onomasticón.

La *Ontología* es el elemento central de FunGramKB. Comprende tres tipos generales de conceptos (llamados metaconceptos y señalados con el símbolo (#)): entidades, (# ENTITY), eventos (# EVENT) y cualidades (# QUALITY). Estos tres conceptos organizan la dimensión cognitiva de los nombres, verbos y adjetivos, respectivamente (Periñán Pascual y Arcas Túnez, 2007). Estas tres dimensiones conceptuales se relacionan entre sí a través de los llamados *postulados de significado* (*Meaning Postulates*, MPs), que son constructos conceptuales que definen la realidad. El siguiente nivel bajo los metaconceptos se compone de los conceptos básicos, precedidos del símbolo (+), y por debajo se encuentran los conceptos terminales, identificados mediante el símbolo (\$), teniendo todo ellos un sufijo numérico al final (_00, _01, etc.). A modo de aclaración, podemos observar la figura que ilustra la esta jerarquía de unidades conceptuales en Periñán Pascual y Arcas Túnez (2007a) para un ejemplo concreto, el concepto terminal \$FOOTBALL_00:

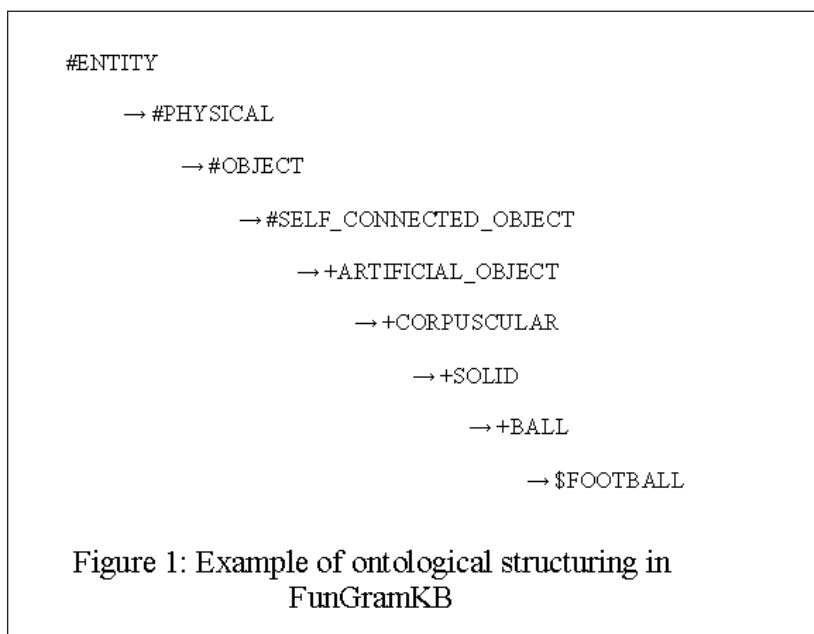


Figura 2. Ejemplo de estructuración ontológica en FunGramKB (Periñán Pascual y Arcas Túnez, 2007:199).

El *Cognición* es el módulo que almacena conocimiento procedimental en forma de secuencias temporales o guiones de situaciones cotidianas (por ejemplo, ir al cine, visitar un museo, ir a un restaurante, etc.), y que está basado particularmente en el modelo temporal de Allen (Allen, 1983; Allen y Ferguson, 1994). En FunGramKB, se estructura como una serie de predicaciones dentro de un marco lineal temporal. Un guión, por tanto, comprende varias predicaciones, y cada una de ellas se entiende como un evento E, tratado como un intervalo entre un par de puntos temporales: el punto temporal *i*, que es el inicio de la acción, y el punto temporal *t*, que es su fin. Así como Allen había previsto la posibilidad de diversas relaciones entre estos eventos (simultaneidad, posterioridad o solapamiento, entre otros), también quedan éstas reflejadas en FunGramKB (Periñán Pascual y Arcas Túnez, 2010b; Periñán Pascual, 2012).

Finalmente, el *Onomasticón* está formado por aquellas entidades que componen nuestro conocimiento enciclopédico o histórico, por ejemplo personas, lugares, acontecimientos, etc., en forma de bio-estructuras. La naturaleza y características de este módulo serán desarrolladas manera algo más extensa en la siguiente sección, con objeto de poder mostrar más adelante cómo el conocimiento cultural almacenado en el Onomasticón de FunGramKB sirve como apoyo para resolver ejemplos concretos de fenómenos anafóricos, que requieren algo más que mero conocimiento lingüístico (léxico o sintáctico) para ser desambiguados con éxito.

1.3. El Onomasticón de FunGramKB

El Onomasticón de FunGramKB está encuadrado dentro del nivel conceptual de la base de conocimiento, junto a la Ontología y al Cognición. El Onomasticón, como se ha reseñado con anterioridad, almacena la información relativa a las instancias de entidades y eventos, en forma de bio-estructuras. Con la finalidad de comprender qué es a lo que se apunta de manera precisa mediante la expresión *bio-estructura*, es necesario mencionar el carácter prototípico del conocimiento almacenado en los módulos conceptuales que son la Ontología y el Cognición.

Dada la prototipicidad que poseen, llamamos a las estructuras que forman la Ontología y el Cognición *proto-estructuras*, puesto que se deben a la generalidad de un concepto y no a la particularidad. No obstante, a pesar de reflejar esta generalidad, en el marco de una situación prototípica, cuando cabe la posibilidad de que alguna de sus características pueda variar en algún momento hipotético del tiempo, se utiliza la etiqueta de *rebatible*, lo que permitirá una herencia denominada *no-monotónica*. Por ejemplo, si hablamos de un pájaro, una característica prototípica de este tipo de animales es “puede volar” pero, puesto que es posible que exista un pájaro que, a pesar de serlo, no sea capaz de volar (p.ej. un pingüino o un avestruz), el rasgo “puede volar” será entonces una característica rebatible.

Por el contrario, el Onomasticón refleja la situación opuesta, donde la prototipicidad deja paso a la especificidad, y puesto que se trata de entidades reales existentes, se utiliza la etiqueta de *bio* para nombrar a estas estructuras.

De este modo, mientras que un concepto de la Ontología es +SONG_00, una entidad del Onomasticón es %HEY_JUDE_00⁴.

Además del parámetro de la prototipicidad, los esquemas conceptuales en FunGramKB también se clasifican conforme a otro parámetro: la temporalidad (Periñán Pascual y Arcas Túnez, 2010b). Teniendo en cuenta esta característica, el conocimiento dentro de los esquemas conceptuales puede presentarse de manera temporal o atemporal. Si se hace de manera temporal, significará que ese conocimiento se presenta dentro de un marco de tiempo (lo que se denominan *macroestructuras*). Por ejemplo, la biografía de una persona (una persona que sea una entidad del Onomasticón) o un guión de los reflejados en el Cognición poseerían la propiedad de ser estructuras temporales. Por otro lado, cabe la posibilidad de que las estructuras conceptuales representen el conocimiento de manera atemporal (*microestructuras*), como sucedería con una característica aislada de una entidad del Onomasticón o de la Ontología. La convergencia de estos dos parámetros, prototipicidad y temporalidad, resulta en la creación de una tipología de estructuras conceptuales compuesta por proto-microestructuras, proto-macroestructuras, bio-microestructuras y bio-macroestructuras. Si se coloca en una matriz esta combinación de parámetros, se obtendrá la siguiente tabla que lo ilustra:

		TEMPORALIDAD	
		-	+
PROTOTIPICIDAD	+	Proto-microestructura (Postulado de significado)	Proto-macroestructura (Guión)
	-	Bio-microestructura (Retrato)	Bio-macroestructura (Historia)

Tabla 1. Tipología de esquemas conceptuales en FunGramKB (Periñán Pascual y Arcas Túnez, 2010b).

Además de estar organizadas de la manera arriba ilustrada, estas cuatro estructuras se integran en la base FunGramKB, lo que puede observarse representado de manera gráfica en la figura siguiente:

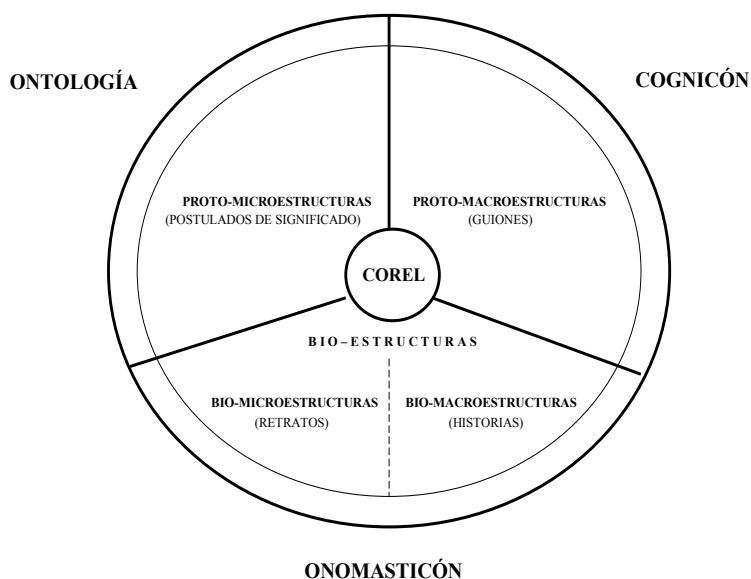


Figura 3. El Planeta Cognitivo (Periñán Pascual y Mairal Usón, 2010:15)

⁴ Nótese el símbolo (%) utilizado para identificar las entidades del Onomasticón, diferente de los símbolos (#), (+) o (\$), que identifican a las entidades pertenecientes a la Ontología como metaconceptos, conceptos básicos o conceptos terminales, respectivamente.

Esta figura muestra la necesaria interrelación entre los tres módulos del nivel conceptual de FunGramKB a través de un mismo lenguaje de representación, denominado COREL⁵, ya que, al igual que ocurre con los componentes de la memoria humana a largo plazo, un sistema de PLN que persiga permitir con éxito el razonamiento ha de prever que los componentes que lo forman puedan relacionarse entre ellos a través de un mismo lenguaje (Periñán Pascual y Arcas Túnez, 2010b; Periñán Pascual y Mairal Usón, 2010).

La Tabla 1 y la Figura 3 arriba mostradas sirven de contexto a la hora de describir el Onomasticón y los elementos que éste alberga. Así pues, si observamos una entidad como, por ejemplo, *Elvis Presley* (%ELVIS_PRESLEY_00, en notación COREL), dependiendo del tipo de característica que se desee reflejar, se necesitará utilizar un tipo u otro de estructura conceptual. Por tanto, si deseamos describir la profesión de cantante de Elvis, será preciso emplear una bio-microestructura, ya que dicha propiedad puede encuadrarse como uno de los muchos rasgos que, de manera relativamente aislada del resto de características, describen a Elvis en un momento concreto del tiempo, como serían, entre otras muchas, la profesión de actor, su color de pelo o su estatura (cf. Periñán Pascual y Arcas Túnez, 2010b). Es por ello que este tipo de microestructuras en el Onomasticón se denominan “retratos” (*snapshots* en su versión en inglés), por el hecho de asemejarse a una fotografía tomada en un momento particular. Sin embargo, si se desea describir la biografía de Elvis, es necesario hacer uso de una bio-macroestructura, las llamadas “historias” (*stories*) por obvias razones: contienen elementos que han de ser integrados dentro de un esquema temporal determinado (p. ej., la fecha de nacimiento o fallecimiento del cantante), ya que se trata de elementos cuya existencia se encuadra en un marco temporal establecido y cuyo orden no es posible alterar.

Otro ejemplo al que se puede aludir es una entidad como el Taj Majal. De nuevo, si perseguimos describir una propiedad como “hecho de mármol” o “mausoleo”, se haría mediante una bio-microestructura (retrato), mientras que el relato de su proceso de construcción habría de realizarse mediante una historia (bio-macroestructura). Esto se puede ilustrar en lenguaje COREL a través del siguiente ejemplo, originalmente sugerido en Periñán Pascual y Arcas Túnez (2010b) acerca de la entidad “%TAH_MAHAL_00”. El equivalente en lenguaje natural de las predicaciones representadas se muestra a continuación de cada una de ellas:

+(e1: +BE_02 (x1: %TAH_MAHAL_00)Theme (x2: %INDIA_00)Location)
El Taj Majal está ubicado en India.

*(e2: +BE_01 (x1)Theme (x3: +WHITE_00 & \$MARBLE_00)Attribute)
Su principal material es el mármol blanco.

*(e3: +COMPRISE_00 (x1)Theme (x4: 1 \$DOME_00 & 4+TOWER_00)Referent)
El Taj Majal tiene una cúpula principal y cuatro torres.

Estas tres predicaciones acerca del Taj Majal pertenecen a la categoría “retrato”, ya que las tres enumeran rasgos atemporales de la entidad. Sin embargo, apreciamos que sólo la primera de ellas (e1), porta el símbolo (+) al frente, lo que indica que tan sólo esa predicación es estricta. Las otras dos predicaciones son rebatibles, lo que se indica en lenguaje COREL mediante el símbolo (*), ya que si el Taj Majal encontrara destruida una de sus torres o la cúpula principal por cualquier circunstancia, o si alguna de éstas se sustituyera por un elemento hecho de otro material diferente del mármol blanco, el Taj Majal seguiría considerándose la misma entidad. No obstante, si el Taj Majal no se encontrase en la India, entonces ya no se trataría de la misma entidad, sino de una reproducción del mausoleo en otra ubicación.

Por otro lado, y continuando con el mismo ejemplo, es posible también representar características temporales del Taj Majal. Es el caso de las siguientes predicaciones y su equivalente en lenguaje natural a continuación (cf. Periñán Pascual y Arcas Túnez, 2010b):

+(e1: past +BUILD_00 (x1)Theme (x2: %TAH_MAHAL_00)Referent (f1: 1633)Time)
El Taj Majal fue construido en 1633.

+(e2: past +BE_00 (x2)Theme (x3:%WORLD_HERITAGE_SITE_00)Referent (f2: 1983)Time)
El Taj Majal se convirtió en Patrimonio de la Humanidad por la UNESCO en 1983.

Los conceptos representados arriba constituyen historias, ya que, además de ser predicaciones estrictas, indicado por el signo (+) frente a cada una de ellas (no es posible cambiar eventos sucedidos en el pasado, como el año de construcción, ni aquel en el que fue declarado Patrimonio de la Humanidad), pertenecen a un esquema temporal en el que están encuadradas de una manera precisa y determinada. Por consiguiente, se observa a través de los ejemplos anteriormente mencionados cómo una misma entidad puede compartir retratos e historias que representen rasgos conceptuales que la definan.

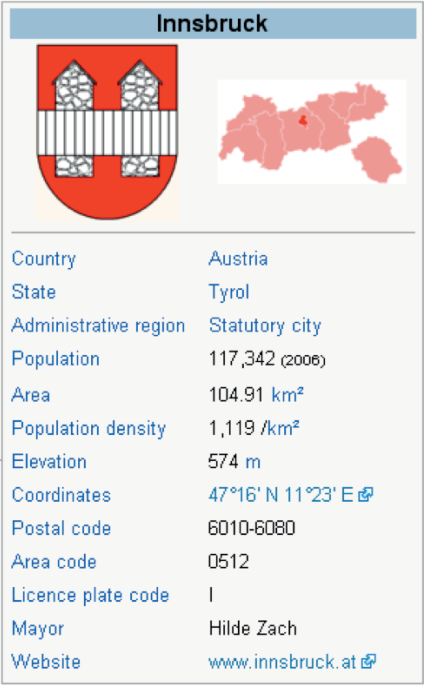
⁵ Conceptual Representation Language, interlengua utilizada como lenguaje de representación a través de todos los módulos conceptuales de FunGramKB.

Otro de los principales rasgos distintivos del Onomasticón frente a Ontología y Cognición es que la manera de poblar este módulo no es enteramente manual, como sucede en los otros dos módulos, sino que se realiza de manera semi-automática, cuyo procedimiento está descrito en Perrián Pascual y Carrión Varela (2011). Mediante esta metodología semi-automática se posibilita que, tras la elaboración manual de plantillas de proyección (denominadas *reglas*), la inserción de éstas en la herramienta FunGramKB permita el trasvase automático de información a FunGramKB procedente de otras fuentes, como es principalmente DBpedia⁶ (Auer *et al.*, 2007; Bizer *et al.*, 2009). Este procedimiento de población semi-automática permite que la gran profusión de datos contenida en DBpedia, cuyo principal origen de importación de datos es Wikipedia⁷, sea transferida a FunGramKB y actualizada a la vez que lo hacen las fuentes que la alimentan. La principal fuente de información que DBpedia toma de Wikipedia para poblar su base de conocimiento son los denominados *info-boxes*, elementos incluidos en los artículos de Wikipedia y cuyo ejemplo se ilustra a continuación:

```

{{Infobox Town AT |
  name = Innsbruck |
  image_coa = InnsbruckWappen.png |
  image_map = Karte-tirol-I.png |
  state = [[Tyrol]] |
  regbzkg = [[Statutory city]] |
  population = 117,342 |
  population_as_of = 2006 |
  pop_dens = 1,119 |
  area = 104.91 |
  elevation = 574 |
  lat_deg = 47 |
  lat_min = 16 |
  lat_hem = N |
  lon_deg = 11 |
  lon_min = 23 |
  lon_hem = E |
  postal_code = 6010-6080 |
  area_code = 0512 |
  licence = I |
  mayor = Hilde Zach |
  website = [http://innsbruck.at] |
}}

```



Country	Austria
State	Tyrol
Administrative region	Statutory city
Population	117,342 (2006)
Area	104.91 km ²
Population density	1,119 /km ²
Elevation	574 m
Coordinates	47°16′ N 11°23′ E﻿ⓘ
Postal code	6010-6080
Area code	0512
Licence plate code	I
Mayor	Hilde Zach
Website	www.innsbruck.at ⓘ

Ilustración 1. Info-box de Wikipedia (Perrián Pascual y Carrión Varela, 2011:93).

Mediante la inclusión en FunGramKB de los contenidos que cada *info-box* proporciona para cada entidad, se compone una estructura conceptual más completa y que permite resolver ejemplos de fenómenos anafóricos complejos, como se describirá a continuación.

2. LA ANÁFORA

Uno de los fenómenos lingüísticos en los que más se ha trabajado a nivel de PLN y cuya resolución ha sido buscada por numerosos proyectos (p. ej., en Mitkov, 2002 se puede encontrar un exhaustivo repaso a dicho propósito a través de la historia) es la anáfora.

La anáfora es una forma de presuposición que consiste en apuntar a un referente o elemento previo mencionado (Halliday y Hasan, 1976), lo que contribuye en gran manera a la cohesión de un texto, especialmente la llamada anáfora endofórica (cuyo antecedente se encuentra dentro del mismo texto), ya que permite la continuación lógica y coherente de una secuencia textual, en lengua escrita u oral, sin necesidad de reiterar continuamente las mismas expresiones lingüísticas o lexicalizaciones para aludir a un mismo concepto o referente. Según resumen Fan *et al.* (2005), existen tres elementos clave en la construcción de una anáfora: el antecedente o anclaje, la expresión que se refiere al antecedente y el enlace o *link*:

The object that is being referred to is called the anchor or the antecedent, the expression that refers to the antecedent is called the referring expression, and the association between the referring expression and the anchor is called the link. (Fan *et al.*, 2005: 153).

⁶ <http://dbpedia.org/About>

⁷ http://en.wikipedia.org/wiki/Main_Page

Para la ilustración de este fenómeno en puridad, así como para mostrar también el tratamiento y resolución que FunGramKB efectúa de otros fenómenos lingüísticos relacionados, en este artículo nos referiremos al término *anáfora* en su concepción más amplia. Esto incluye a otros fenómenos relacionados como, por ejemplo, la catáfora, puesto que el objetivo principal de los ejemplos y descripciones presentados más adelante es vincular las cadenas de correferencia existentes en el lenguaje natural, tanto las que hacen alusión a elementos previamente mencionados, como sucede en la anáfora, así como aquellas cadenas correferenciales materializadas mediante la mención posterior de la entidad nombrada, lo que conocemos como catáfora.

2.1. Resolución de antecedentes anafóricos de entidades nombradas

Si bien para la mayoría de los hablantes nativos de una lengua natural es algo relativamente sencillo el identificar y resolver el antecedente al que hace alusión una expresión anafórica, no siempre sucede de este modo para las máquinas que procesan el lenguaje natural. Con objeto de ilustrar este fenómeno, observemos los siguientes ejemplos en los cuales se han destacado las alusiones a entidades nombradas (antecedentes) y sus correspondientes anáforas:

- (i) Good morning from Hollywood. You know, here at the famous Grauman's Chinese Theater, they have handprints and footprints from all sorts of famous people, from Jimmy Stewart to the cast of Harry Potter. And this will be the latest, **Michael Jackson**, the imprint of **his** famous sequined glove over there, **his** footprints. And then the handprints of **his** three children who, just hours ago, took steps to ensure **their father's** Hollywood immortality.⁸
- (ii) The Nation has since denied any "official business or professional relationship" with Jackson, and yet several of their members, sober-faced and in business suits, lined up outside the Santa Maria courthouse at the singer's arraignment last January. More than a thousand Jackson supporters thronged the streets in front of the courthouse, holding up signs and cheering as they strained for a glimpse of Michael, who was accompanied that morning by his mother, father, brother Jermaine and sister Janet. Suddenly a roar went up. There was the **King of Pop** on the roof of a vehicle waving his arms and blowing kisses at the screaming crowd.⁹

En (i), se observa que la referencia al fallecido cantante *Michael Jackson* se replica mediante la utilización en tres ocasiones del determinante posesivo *his*, lo que facilita la cohesión textual y permite identificar, aun en el supuesto caso de ignorar quién es *Michael Jackson*, que *his* se refiere en las tres ocasiones al mismo referente o entidad masculina. Además, la mención posterior como *their father* se debe unir al referente *his three children*, que a nivel conceptual o semántico queda vinculado a *their father*, lo que hace posible interpretar el texto de manera coherente por parte del receptor del mensaje.

En caso de ser interpretado por una máquina, concretamente por la herramienta FunGramKB, la interpretación de las expresiones *his* en relación a *their father* contenidas en (i) podría ser resuelta sin mayor problemática, gracias a la carga conceptual y morfológica que poseen los diferentes módulos de la base de conocimiento y que permite, por medio de su razonador, vincular los conceptos entre sí¹⁰.

Sin embargo, para identificar el sexo de la entidad *Michael Jackson* y, de este modo, el correspondiente género gramatical que permitiría la vinculación anafórica entre la entidad nombrada, el determinante posesivo *his* y el posterior *their father*, resulta de gran importancia el módulo Onomástico de FunGramKB. En este ejemplo concreto, gracias a la importación de la información contenida en Wikipedia, es posible saber que la entidad *Michael Jackson* es un cantante de sexo masculino, como se ilustra a continuación:

⁸ For January 27, 2012, CBS 2012 (120127). CBS_ThisMorning. Obtenido a través del corpus COCA, Davies (2008-).

⁹ Cheo Hodari Coker, MAN IN THE MIRROR, 2004 (Apr) Vol. 34, Iss. 12; pg. 186, 4 pgs, Essence. Obtenido a través del corpus COCA, Davies (2008-).

¹⁰ Por ejemplo, gracias al conocimiento conceptual almacenado en la Ontología, el razonador de FunGramKB sabe que un padre es una persona de género masculino que tiene descendencia.

The screenshot shows the Wikipedia page for Michael Jackson. At the top, there's a navigation bar with 'Article', 'Talk', 'Read', 'View source', and 'View history'. The main heading is 'Michael Jackson' with a subtext 'From Wikipedia, the free encyclopedia'. Below this, there's a disambiguation note: 'For other people named Michael Jackson, see Michael Jackson (disambiguation)'. The main text begins with 'Michael Joseph Jackson^{[1][2]} (August 29, 1958 – June 25, 2009) was an American singer-songwriter, dancer, businessman and philanthropist. Often referred to by the honorific nickname "King of Pop", or by his initials MJ^[3] Jackson is recognized as the most successful entertainer of all time by Guinness World Records. His contributions to music, dance, and fashion, along with his publicized personal life, made him a global figure in popular culture for over four decades.' To the right of the text is a photograph of Michael Jackson performing on stage, wearing a dark jacket with a star on the sleeve. Below the photo is a caption: 'Jackson performs on his *Bad* world tour in 1988'. Underneath the photo is a 'Background information' table with fields for 'Birth name', 'Also known as', and 'Born'.

Ilustración 2. La entidad "Michael Jackson" en Wikipedia¹¹

En relación al ejemplo (ii), éste pone de manifiesto de manera más explícita la necesidad e importancia de la existencia de un módulo en FunGramKB que se ocupe de nutrir la base de conocimiento con información acerca de entidades nombradas. De otro modo, la alusión a *King of Pop* que aparece en dicho texto no podría ser resuelta de manera exitosa por parte de la máquina, la cual, basándose tan sólo en coincidencias léxico-gramaticales, tendría dificultades en determinar qué entidad nombrada masculina mencionada anteriormente en el texto (*father, brother Jermaine, Michael*) desempeña el papel de antecedente de la expresión *King of Pop*. Gracias a la información cultural y enciclopédica incluida en el Onomasticón e importada de Wikipedia, sabemos que *King of Pop* se refiere a *Michael Jackson*, como evidencia el artículo de Wikipedia mostrado en la Ilustración 2.

Otro ejemplo que mostraría la capacidad de FunGramKB para desambiguar cadenas correferenciales lo podemos obtener de la obra de Mitkov (2002), donde se menciona una cadena correferencial (expresiones que tienen el mismo referente a lo largo del mismo extracto textual):

- (iii) *Sophia Loren* says *she* will always be grateful to Bono. *The actress* revealed that the U2 singer helped *her* calm down when *she* became scared by a thunderstorm while travelling on a plane. (Mitkov, 2002:5).

En (iii) podemos observar cómo se efectúa una cadena de correferencias entre *Sophia Loren*, *she* y *the actress*, tal como destaca el autor mediante el empleo de la letra cursiva. Por otro lado, encontramos que entre "Bono" y "U2 singer" también existe una correferencia. Sin embargo, en caso de que una máquina tuviera que decidir cuáles son las cadenas de correferencia en este extracto, podría existir ambigüedad a la hora de decidir si *the actress* hace referencia a *Sofía Loren* o a "Bono", al igual que la mención a "the U2 singer". No obstante, puesto que dicha información se encuentra dentro de los datos que Wikipedia importa al Onomasticón de FunGramKB, de nuevo quedaría salvaguardada la correcta interpretación referencial.

2.2. Resolución de la anáfora indirecta

En ocasiones, la coherencia de un extracto textual va más allá de una mera cadena correferencial existente en la misma oración o extracto, para desplazarse a referentes más alejados, incluso más allá del emisor, el receptor o los elementos contextuales del momento de la emisión del mensaje. Este fenómeno es lo que se denomina anáfora indirecta. Según Mitkov (2002:15), "*Indirect anaphora arises when a reference becomes part of the hearer's or reader's knowledge indirectly rather than by direct mention*", lo que significa que la conexión de una expresión anafórica con su antecedente resulta más complicada, al no estar basada en una identificación sintácticamente detectable, sino que va más allá, precisando de conocimiento general del mundo. Otros autores se refieren a fenómenos relacionados mediante diferentes denominaciones, como por ejemplo Eckert y Strube (citados en Palomar *et al.*, 2000), que lo denominan *anáfora abstracta*, si bien es cierto que este término lo aplican de una manera más general a cualquier antecedente que no sea un sintagma nominal (Palomar *et al.*, 2000: 206): "*According to Eckert and Strube (1999), if the antecedent is a noun phrase then the anaphora is classified as individual anaphora, otherwise, the anaphora is classified as abstract anaphora.*" Un nuevo término que podemos relacionar también es el concepto de *bridging* (Palomar *et al.*, 2000; Fan *et al.*, 2005), referido a las relaciones

¹¹ http://en.wikipedia.org/wiki/Michael_Jackson [fecha de consulta 15/10/2013].

(excluyendo la de identidad) que existen entre una anáfora y su antecedente, lo que nos permite encuadrar aquí relaciones como la hiponimia o hiperonimia:

Clark (1977) called bridging descriptions to definite descriptions that either have an antecedent denoting the same discourse entity, but using a different head noun (synonym, hypernym or hyponym) or are related by other relation than identity. (Palomar et al., 2000: 207).

Debido a la gran complejidad que entraña la resolución de la anáfora indirecta, diversos autores (Mitkov, 2002; Fan et al., 2005) concluyen a este respecto lo que se resume en Mitkov (2002:34): “Therefore anaphors requiring real-world knowledge for their resolution stand the least chance of being resolved successfully”, lo que parece confirmar que, si en muchos casos este tipo de anáforas pueden resultar complicadas de resolver incluso para un humano que no posea ciertos conocimientos culturales, la resolución por parte de la máquina puede tornarse una quimera. Precisamente por esto, es en estos casos donde el Onomasticón de FunGramKB puede resultar de especial ayuda para conferir sentido a las referencias expresadas, ya que la integración de conocimiento enciclopédico en este módulo supone un enriquecimiento cultural de la herramienta, lo que permite la resolución de casos de anáfora indirecta, como veremos a continuación.

Observemos el siguiente ejemplo:

(iv) When *Take That* broke up, the critics gave *Robbie Williams* no chance of success. (Mitkov, 2002:15).

En este caso, la información acerca de lo que es la entidad *Take That* es necesaria para comprender dónde está la correlación o coherencia semántica de aludir a otra entidad no mencionada previamente, *Robbie Williams*, ya que a simple vista no parece establecerse ninguna cadena correferencial. Sin embargo, la motivación de mencionar a la entidad *Robbie Williams* en este ejemplo responde a su condición de integrante de la banda musical inglesa *Take That*, una información que Wikipedia (y, por ende, el Onomasticón de FunGramKB) sí es capaz de aportar acerca de la banda y sus integrantes:

The image shows a screenshot of the Wikipedia article for 'Take That'. The article title is 'Take That' and it is described as an English pop group. The text mentions members Gary Barlow, Howard Donald, Jason Orange, Mark Owen, and Robbie Williams. It details their career, including their debut album 'Partners in Crime' in 1992, their reunion in 2006, and their 2011 album 'Progress'. A sidebar on the right provides background information such as their origin (Manchester, England), genres (Pop, pop rock, dance), and active years (1990–96, 2005–present). A photo of the band performing is also visible.

Ilustración 3. La entidad “Take That” en Wikipedia¹²

Otra instancia más compleja de cómo FunGramKB es capaz de resolver este problema se puede observar en el ejemplo que mostramos en (v). Para poder proporcionar una respuesta exitosa a la desambiguación anafórica del extracto, es necesario que la herramienta posea información relativa a eventos de corte histórico, lo que va más allá de un conocimiento básico de entidades nombradas individuales, para referirse a eventos complejos de la historia, conocidos por los hablantes medios de una comunidad pero ignorados, en su gran mayoría, por las máquinas. Observemos la siguiente expresión destacada:

(v) *It is possible that some human should have climbed the World Trade Center Towers without ropes before they were destroyed.* (King, 2013).

¹² http://en.wikipedia.org/wiki/Take_That [fecha de consulta 21/10/2013].

En este caso, para que la máquina sea capaz de efectuar la resolución anafórica del antecedente del pronombre *they*, es necesario que ésta posea conocimiento acerca de que las torres del World Trade Center fueron destruidas en una fecha precisa, ya que los humanos a los que se menciona en el mismo ejemplo, tanto sintáctica como conceptualmente, podrían ser un potencial antecedente¹³. No obstante, el Onomasticón proporciona el vínculo conceptual con el conocimiento enciclopédico necesario para desambiguar el antecedente del pronombre *they*, como se ilustra en la Figura 3, que muestra una representación gráfica de la conexión conceptual entre la información del *info-box* de Wikipedia para la entidad “World Trade Center Towers”¹⁴, la fecha de destrucción de dicha entidad y la propiedad “*destructionDate*” del Onomasticón:

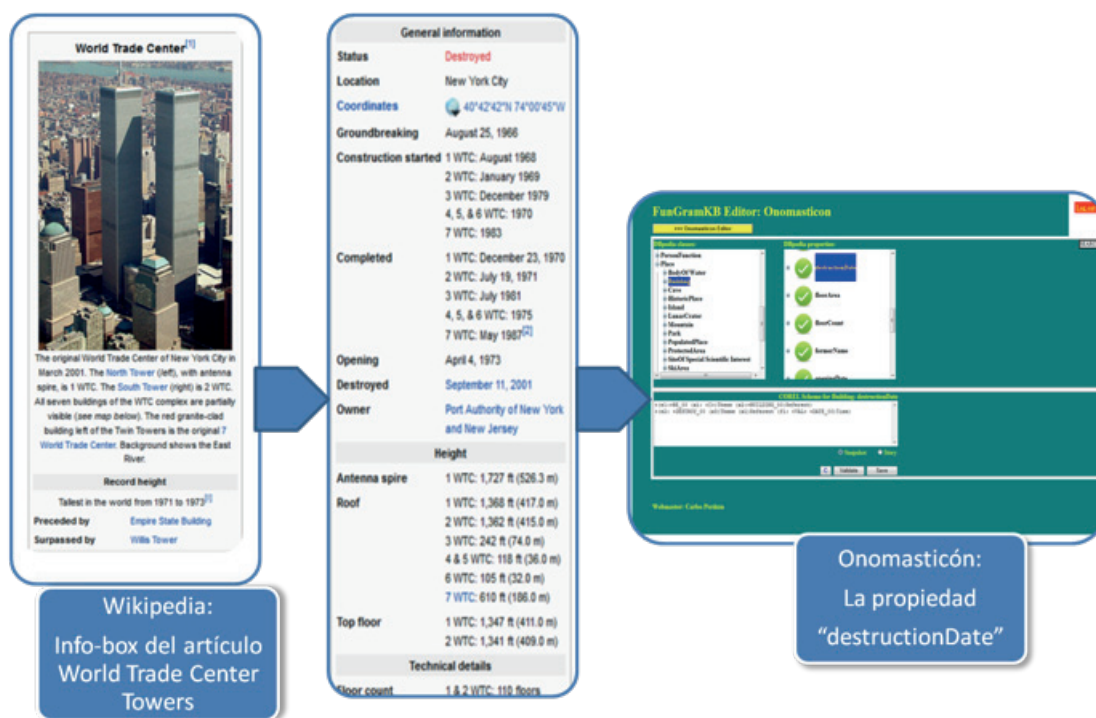


Figura 4. Vínculo conceptual de las propiedades de la entidad “World Trade Center Towers” en Wikipedia y el Onomasticón de FunGramKB.

En la Figura 4 observamos la siguiente conexión conceptual: la información que Wikipedia aporta acerca del hecho histórico de la destrucción de las torres del World Trade Center se basa tanto en la categorización de estado de la entidad (“*Status: Destroyed*”) como en la asignación de una fecha concreta a dicho acontecimiento (“*September 11, 2001*”). La conexión de esta información con FunGramKB la encontramos a través del Onomasticón, a través de la propiedad “*destructionDate*”. Esta propiedad determina la fecha en la que cualquier entidad catalogada como “Building” (la entidad World Trade Center Towers lo es) ha sido destruida (si lo ha sido). Por tanto, este vínculo conceptual permite que, en el ejemplo (v), se resuelva de una manera más clara y menos ambigua que el antecedente del pronombre *they* es la entidad World Trade Center Towers, lo que pone de manifiesto la capacidad de la base de conocimiento FunGramKB para actuar de apoyo a la hora de resolver de manera más exitosa cuestiones y problemas derivados del PLN.

3. CONCLUSIONES

En el presente artículo se ha tratado de mostrar cómo FunGramKB y la inclusión de conocimiento cultural a través del Onomasticón posibilitan la ampliación del potencial de dicha herramienta, con objeto de ayudar a desambiguar fenómenos anafóricos en los que se requiere conocimiento cultural o enciclopédico. Puesto que la concepción de FunGramKB como herramienta multipropósito es una premisa sobre la cual se basa nuestra labor de investigación presente y futura, en los trabajos actualmente en desarrollo se incluye la aplicación del Onomasticón de FunGramKB a la resolución de otros problemas lingüísticos que acucian a los sistemas de PLN, como son la referencia y correferencia, metáfora y metonimia. Dicho trabajo encontrará su publicación en la tesis doctoral actualmente en elaboración por parte de la autora del presente artículo.

¹³ Puesto que los humanos, según la información conceptual contenida en la Ontología de FunGramKB, son entidades que también pueden ser destruidas.

¹⁴ http://en.wikipedia.org/wiki/World_Trade_Center_Towers, fecha de consulta [23/10/2013].

REFERENCIAS

- Allen, J.F. (1983). "Maintaining knowledge about temporal intervals", *Communications of the ACM*, 26/11: 832-843. doi:10.1145/182.358434
- Allen, J.F. y Ferguson, G. (1994). "Actions and events in temporal logic", *Journal of Logic and Computation*, 4/5: 531- 579. doi:10.1093/logcom/4.5.531
- Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., e Ives, Z. (2007). "DBpedia: A Nucleus for a Web of Open Data", en Aberer *et al.* (Eds.), *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007*. Berlin/Heidelberg: Springer. Disponible en <http://www.informatik.uni-leipzig.de/~auer/publication/dbpedia.pdf> [fecha de consulta: 14/10/2013].
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. y Hellmann, S. (2009). "DBpedia – A Crystallization Point for the Web of Data", *Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7*: 154-165. Disponible en http://jens-lehmann.org/files/2009/dbpedia_jws.pdf [fecha de consulta: 14/10/2013].
- Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Disponible en <http://corpus.byu.edu/coca> [fecha de consulta: 13/11/2013].
- Fan, J., Barker, K. y Porter, B. (2005). "Indirect anaphora resolution as semantic path search", en *Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05)*. ACM, New York, NY, USA, pp. 153-160. Disponible en: <http://doi.acm.org/10.1145/1088622.1088650> [fecha de consulta: 28/11/2013].
- Halliday, M.A.K. y Hasan, R. (1976). *Cohesion in English*. London: Longman.
- King, J.C. (2013). Anaphora, "*The Stanford Encyclopedia of Philosophy*" (Summer 2013 Edition), Edward N. Zalta (ed.), Disponible en: <http://plato.stanford.edu/archives/sum2013/entries/anaphora/> [fecha de consulta: 23/10/2013].
- Mairal Usón, R. (2012). "La arquitectura de una base de conocimiento léxico conceptual: implicaciones lingüísticas", en M. Giammatteo, L. Ferrari y H. Albano (eds.). *Léxico y Sintaxis*. Volumen temático de la serie editada por la Sociedad Argentina de Lingüística. Mendoza: Editorial FFyL, UNCuyo y SAL, pp. 183-210. Disponible en <http://ffyl.uncu.edu.ar/spip.php?article3638> [fecha de consulta: 13/11/2013].
- Mairal Usón, R., Periñán Pascual, C. y Pérez Cabello de Alba, M. B. (2012). "La representación léxica. Hacia un enfoque ontológico", en R. Mairal Usón, L. Guerrero y C. González (eds.) *El funcionalismo en la teoría lingüística. La Gramática del Papel y la Referencia. Introducción, avances y aplicaciones*. Madrid: Akal, pp. 85-102.
- Mitkov, R. (2002). *Anaphora Resolution*. Pearson Education Limited. Great Britain: Longman.
- Palomar Sanz, M., Saiz Noeda, M., Muñoz Guillena, R., Suárez Cueto, A. y Martínez Barco, P. (2000). "PHORA: a system to solve the anaphora in Spanish", en *Proceedings of Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000)*. Lancaster: University of Lancaster, 2000, pp. 206-211.
- Periñán Pascual, C. (2012). "The situated common-sense knowledge in FunGramKB", *Review of Cognitive Linguistics*, 10/1: 184-214. doi:10.1075/rcl.10.1.06per
- Periñán Pascual, C. y Arcas Túnez, F. (2004). "Meaning postulates in a lexico-conceptual knowledge base", en *Proceedings of the 15th International Workshop on Databases and Expert Systems Applications*. Los Alamitos (California): IEEE, pp. 38-42.
- Periñán Pascual, C. y Arcas Túnez, F. (2006). "Reusing Computer-oriented Lexica as Foreign-Language Electronic Dictionaries", *Anglogermanica online: Revista electrónica periódica de filología alemana e inglesa, n° 4*, pp. 69-93. Disponible en: <http://anglogermanica.uv.es:8080/Journal/Viewer.aspx?Year=2006&ID=perinan.pdf> [fecha de consulta: 14/10/2013].
- Periñán Pascual, C. y Arcas Túnez, F. (2007). "Cognitive modules of an NLP knowledge base for language understanding", *Procesamiento del Lenguaje Natural*, 39: 197-204.
- Periñán Pascual, C. y Arcas Túnez, F. (2008). "A cognitive approach to qualities for NLP", *Procesamiento del Lenguaje Natural*, 41: 137-144.
- Periñán Pascual, C. y Arcas Túnez, F. (2010a), "Ontological commitments in FunGramKB", *Procesamiento del Lenguaje Natural*, 44: 27-34.
- Periñán Pascual, C. y Arcas Túnez, F. (2010b). "The Architecture of FunGramKB", en *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Malta, ELRA, pp. 2667-2674.
- Periñán Pascual, C. y Carrión Varela, M.d.I.L. (2011). "FunGramKB y el conocimiento cultural", en *Anglogermanica Online 2011*, pp. 87-105. Disponible en: <http://anglogermanica.uv.es:8080/Journal/Viewer.aspx?Year=2011&ID=pericarrion.pdf> [fecha de consulta: 14/10/2013].
- Periñán Pascual, C. y Mairal Usón, R. (2009). "Bringing Role and Reference Grammar to natural language understanding", *Procesamiento del Lenguaje Natural*, 43: 265-273.

- Periñán Pascual, C. y Mairal Usón, R. (2010). “La gramática de COREL: un lenguaje de representación conceptual”, *Onomázein*, 21: 11-45.
- Periñán Pascual, C. y Mairal Usón, R. (2012). “La dimensión computacional de la Gramática del Papel y la Referencia: la estructura lógica conceptual y su aplicación en el procesamiento del lenguaje natural”, en R. Mairal Usón, L. Guerrero y C. González (eds.). *El funcionalismo en la teoría lingüística. La Gramática del Papel y la Referencia. Introducción, avances y aplicaciones*. Madrid: Akal, pp. 333-348.
- Van Valin, R.D. Jr y Mairal Usón, R. (en prensa). “Interfacing the Lexicon and an Ontology in a Linking Algorithm”, en M. Ángeles Gómez, F. Ruiz de Mendoza y F. González-García (eds.). *Form and Function in Language: Functional, Cognitive and Applied Perspectives. Essays in Honour of Christopher S. Butler*. Amsterdam: John Benjamins.