Introducción a FunGramKB

Carlos Periñán Pascual, Universidad Politécnica de Valencia (España)
Francisco Arcas Túnez, Universidad Católica San Antonio (España)

Índice

Resumen

Los vínculos entre el procesamiento del lenguaje natural y la lingüística teórica no han sido ni tan numerosos ni tan productivos como podríamos esperar. En los últimos veinte años, ha aumentado progresivamente la tendencia a utilizar enfoques estadísticos, los cuales resultan más económicos y rápidos de implementar. No obstante, con el fin de realizar algún avance en el procesamiento semántico y pragmático, se requiere un nuevo paradigma en los sistemas de comprensión del lenguaje, en el cual se conjuguen los resultados investigadores de disciplinas como la ciencia cognitiva, la lingüística y la inteligencia artificial. Con este objetivo, se diseñó e implementó computacionalmente la base de conocimiento FunGramKB. Este artículo presenta brevemente los diversos módulos que configuran los niveles léxico, gramatical y conceptual en esta base de conocimiento, destaca aquellas investigaciones que han influido de manera más determinante en nuestro modelo teórico y además describe el papel de este recurso en un sistema del procesamiento del lenguaje natural.

Palabras clave: FunGramKB, base de conocimiento, procesamiento del lenguaje natural, razonamiento, Gramática del Papel y la Referencia

1 Introducción

Los lingüistas suelen hablar de la lingüística computacional como un área de conocimiento dentro de la lingüística aplicada. En cambio, debido a la posibilidad de desarrollar sistemas de computación que simulen algún aspecto de la capacidad lingüística del ser humano, los informáticos consideran la lingüística computacional como una rama de la inteligencia artificial, al igual que los sistemas expertos o la robótica, en cuyo caso prefieren hablar de procesamiento del lenguaje natural (PLN). De hecho, un sistema del PLN puede ser concebido como un sistema experto, integrando así ambos campos en lo que se conoce como "ingeniería del conocimiento". La investigación en sistemas expertos tiene como finalidad la construcción de aplicaciones inteligentes que puedan resolver problemas complejos en contextos profesionales. En el caso del PLN, estos problemas implican la implementación computacional de alguna destreza lingüística. Cuando equiparamos

un sistema del PLN con un sistema experto, podemos deducir fácilmente que uno de los componentes centrales de nuestra aplicación es la base de conocimiento, ya que, además de la separación modular entre el conocimiento y el resto del sistema, la característica esencial de los sistemas expertos es la representación explícita de dicho conocimiento (Adarraga 1994). Por tanto, el lingüista que desee implicarse en un proyecto del PLN debe convertirse, en mayor o menor grado, en un ingeniero del conocimiento. Dentro de este marco, un nutrido grupo de investigadores están desarrollando FunGramKB, una base de conocimiento que propicia la construcción de sistemas para la comprensión del lenguaje fundamentados en la lingüística teórica y la ciencia cognitiva.

Después de una breve descripción de la arquitectura de FunGramKB (capítulo 2), presentamos las principales influencias teóricas sobre el modelo de nuestra base de conocimiento (capítulo 3). Finalmente, describimos el papel que desempeña FunGramKB en el procesamiento textual (capítulo 4) y el papel del lingüista en un proyecto de ingeniería lingüística (capítulo 5).

2 Definición y arquitectura de FunGramKB

FunGramKB[1] es una base de conocimiento léxico-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 (p.ej. recuperación y extracción de información, traducción automática, sistemas basados en el diálogo, etc) y con diversas lenguas.[2] 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:[3]

Nivel léxico:

(i) El Lexicón almacena la información morfosintáctica de las unidades léxicas.

(ii) El Morficón asiste al analizador y 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 (MLC) (Ruiz de Mendoza y Mairal Usón, 2008; Mairal Usón y 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 (GPR) (Van Valin and 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

[1] www.fungramkb.com

[2] Actualmente, FunGramKB ha sido modelada para poder trabajar con siete lenguas: alemán, búlgaro, catalán, español, francés, inglés e italiano.

[3] Para una información más detallada sobre el conocimiento almacenado en FunGramKB, léanse Mairal Usón y Periñán Pascual (2009) con respecto al nivel léxico, y Periñán Pascual y Arcas Túnez (2004, 2007a, 2008, 2010a, 2010b) y Periñán Pascual y Mairal Usón (2010a, 2010b) sobre el nivel conceptual.

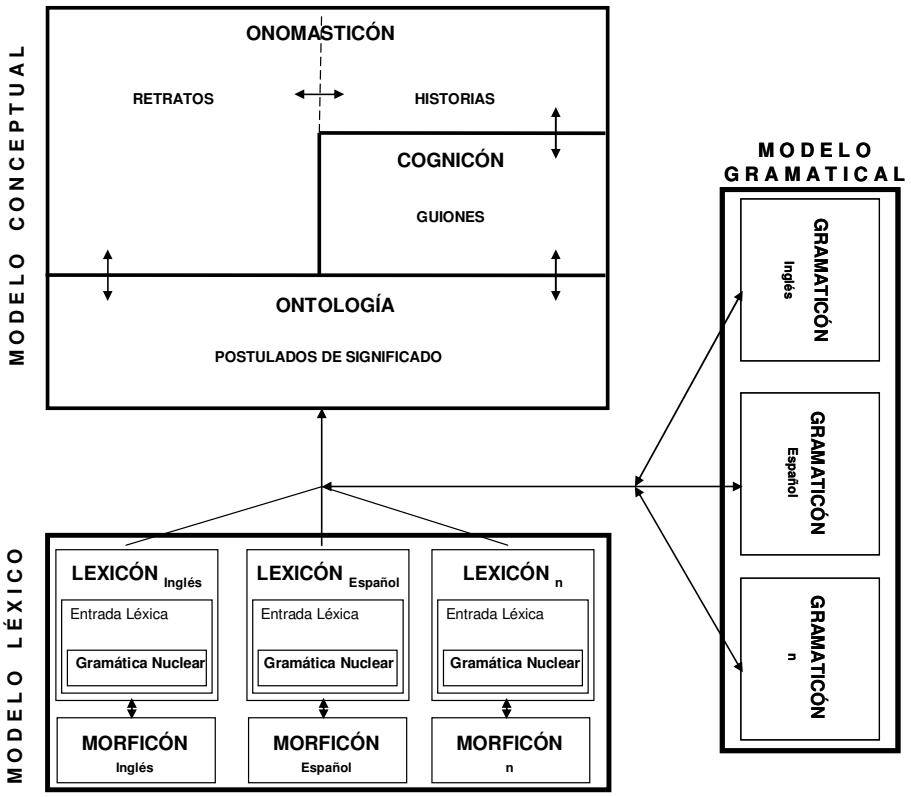
monotónica,[4] consiste en dos tipos de componentes: un módulo de propósito general (i.e. Ontología Nuclear) y varios módulos terminológicos específicos del dominio (i.e. Ontologías Satélites).

(v) El Cognicón almacena el conocimiento procedimental por medio de guiones, i.e. esquemas conceptuales que describen una serie de eventos estereotípicos 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 enciclopédico 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.

Como vemos en la Figura 1, 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 restantes lenguas, pero los ingenieros del conocimiento sólo necesitan construir una Ontología, un Cognicón y un Onomasticón para procesar conceptualmente un texto de entrada.

Figura 1. La arquitectura modular de FunGramKB.



[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 y Arcas Túnez (2010a) describen el tratamiento de este tipo de herencia en el modelo ontológico de FunGramKB.

En este escenario, FunGramKB adopta un enfoque conceptualista, ya que la Ontología se convierte en el pivote de toda la arquitectura de la base de conocimiento. La mayoría de teorías lingüísticas contemporáneas siguen siendo lexicistas, en el sentido de que las unidades léxicas no son consideradas como átomos de la gramática sino como objetos complejos generalmente representados por estructuras de rasgos (Gibbon 2000). En cambio, un enfoque conceptualista, el cual comparte las ventajas del modelo lexicista, es más adecuado para el tratamiento del multilingüismo dentro del marco del PLN.

3 Influencias teóricas sobre el modelo de FunGramKB

3.1 FunGramKB y la ciencia cognitiva

El modelo de "esquema" originado en la psicología cognitiva, e implementado posteriormente en inteligencia artificial, es fundamental para la representación del conocimiento conceptual en FunGramKB. Según este enfoque, un esquema es una representación mental de una entidad o evento, la cual consiste generalmente en un conjunto de expectativas que se van desarrollando a medida que los recuerdos se nutren de experiencias similares. Típicamente los esquemas contienen conocimiento generalizado a partir de las experiencias pasadas, facilitando así la inferencia de información a partir de nuestra percepción del mundo. Las experiencias futuras se interpretan de acuerdo con los patrones construidos a partir de las experiencias pasadas, lo cual alivia la sobrecarga cognitiva. La naturaleza psicológica de los esquemas mentales, i.e. que las personas procesan los intercambios lingüísticos con respecto a un conjunto de expectativas almacenadas en nuestra mente, se fundamenta en el hecho de que las predicciones erróneas surgen generalmente de las correspondencias erróneas entre la información perceptual y el modelo estandarizado del mundo. Por ejemplo, si escuchamos la oración *Javier va a la universidad por las tardes*, podríamos inferir que *Javier* es un estudiante, según el conocimiento estandarizado sobre qué tipo de personas generalmente van a la universidad. Si más tarde escuchamos que *El próximo fin de semana Javier tendrá que corregir más de 500 exámenes*, entonces inferiremos con un mayor grado de certeza que se trata de un profesor.

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 la unidad léxica *novela* implica describir la protoestructura del concepto al que va asignada; en cambio, si deseamos proporcionar información sobre la novela *La Rosa del Viento*, 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 de la biografía de Carlos Ruiz Zafón requiere una macroestructura, mientras que una microestructura es suficiente para describir la profesión de escritor. Cuando combinamos estos dos parámetros, obtenemos la tipología de esquemas conceptuales mostrada en la Tabla 1.

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

Tabla 1. Tipología de los esquemas conceptuales en FunGramKB.

En FunGramKB, todo este conocimiento semántico (p.ej. postulados de significado), procedimental (p.ej. guiones) y episódico (p.ej. bioestructuras) 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).[5] A modo de ilustración, presentamos algunas de las predicaciones correspondientes al retrato (1a) y a la historia (2a) asignados a %TAH_MAHAL_00, cuyos equivalentes al español se presentan en (1b) y (2b) respectivamente.

(1a) +(e1: +BE_02 (x1: %TAH_MAHAL_00)Theme (x2: %INDIA_00)Location)
*(e2: +BE_01 (x1)Theme (x3: +WHITE_00 & $MARBLE_00)Attribute)
*(e3: +COMPRISE_00 (x1)Theme (x4: 1 $DOME_00 & 4 +TOWER_00)Referent)

(1b) El Tah Mahal está en la India.
Su material es mármol blanco.
Tiene una cúpula y cuatro torres.

(2a) +(e1: past +BUILD_00 (x1)Theme (x2: %TAH_MAHAL_00)Referent (f1: 1633)Time)
+(e2: past +BE_00 (x2)Theme (x3: %WORLD_HERITAGE_SITE)Referent (f2: 1983)Time)

(2b) El Tah Mahal fue construido en 1633.
Fue declarado Patrimonio de la Humanidad en 1983.

Desde el enfoque de la ciencia cognitiva, FunGramKB organiza su conocimiento en categorías taxonómicas y categorías derivadas de objetivos. Mientras las categorías

[5] Véase Periñán Pascual y Mairal Usón (2010b) para una descripción detallada del lenguaje de notación de COREL.

taxonómicas van provistas de representaciones independientes del contexto organizadas jerárquicamente a través de un modelo ontológico, las categorías derivadas de objetivos se conceptualizan a través de representaciones que tienen en cuenta la situación de fondo. Barsalou (1991) demostró que los hablantes de una comunidad comparten estructuras de prototipo estables en el tiempo, las cuales desempeñan un papel central en la representación y el procesamiento tanto de las categorías taxonómicas como de las derivadas de objetivos. En este sentido, FunGramKB implementa el enfoque de la "conceptualización situada" (Barsalou 2002) a través de las categorías derivadas de objetivos, las cuales adoptan la forma de macroestructuras conceptuales. Según este enfoque, las personas conceptualizamos una categoría de manera diferente a lo largo de las diferentes situaciones, siendo cada conceptualización relevante para su correspondiente situación. Por ejemplo, nuestra representación prototípica de una silla no es la misma cuando la situamos en una casa, un colegio o una oficina. Este modelo implica que el conocimiento conceptual se organice en una serie de situaciones con el fin de facilitar el procesamiento. En la comprensión lingüística, por ejemplo, las situaciones de fondo ejercen un papel decisivo para la consecución de la tarea. Las personas utilizamos los modelos de situación para representar los significados de los textos. Identificando la situación, se restringen las entidades y los eventos que probablemente ocurran, los cuales constituyen el conocimiento relevante en la situación actual. De hecho, una organización del conocimiento en torno a las situaciones implica que el procesamiento cognitivo se vuelva más tratable. En otras palabras, en lugar de recorrer todos los elementos de la memoria, el sistema se centra exclusivamente en las entidades y los eventos más relevantes para la situación. De esta forma, resulta más fácil recuperar información útil no sólo para la comprensión textual, sino también para la resolución de problemas, el razonamiento y la predicción de las acciones (Yeh y Barsalou 2006).

3.2 FunGramKB y la lingüística teórica

Las teorías lingüísticas pueden ser divididas *grosso modo* bajo dos grandes paradigmas: las teorías basadas en la forma de las expresiones lingüísticas (p.ej. modelos generativos) y aquellas basadas en la función de las expresiones lingüísticas (p.ej. modelos funcionales). Los modelos generativos dominaron el panorama de la lingüística computacional debido principalmente a su explicitud formal. Por el contrario, aunque los modelos funcionales tratan sobre el uso adecuado de la lengua, frecuentemente no logran una formalización detallada que facilite la implementación computacional (Fraser 1991). Desde los años 90, el PLN ha cambiado su foco de atención, yendo desde la formalización manual del conocimiento lingüístico a la aplicación de técnicas de aprendizaje automático sobre extensos córpora. Este giro metodológico se origina por un cambio en el objetivo de la investigación: se ha pasado de la investigación teórica fundamental a las aplicaciones de la ingeniería lingüística, lo cual ha sido propiciado principalmente por la disponibilidad de extensos repositorios de datos. Este cambio de interés ha dado lugar, por ejemplo, a la construcción de bases de datos que adoptan un enfoque relacional en la representación del significado léxico, ya que resulta más fácil indicar las asociaciones entre las unidades léxicas en forma de relaciones semánticas que describir formalmente el contenido conceptual de las unidades léxicas. En cambio, aunque el desarrollo a gran escala de recursos orientados a la semántica profunda requiere mucho más tiempo y esfuerzo, los significados conceptuales presentan un poder expresivo más robusto, además de facilitar la gestión del conocimiento de manera más efectiva (cf. Periñán Pascual y Arcas Túnez, 2007). Además, la mayoría de las bases de datos léxicas no están fundamentadas en sólidas teorías lingüísticas, lo cual no facilita las generalizaciones sintáctico-semánticas que permitan tanto la explicación como la predicción de fenómenos lingüísticos. Evidentemente, es mucho más rápido construir sistemas del

PLN cuando se ignoran las teorías lingüísticas, pero las aplicaciones del PLN que pueden funcionar perfectamente sin la necesidad de basarse en algún modelo lingüístico son "engañosamente inteligentes" (Halvorsen, 1988), ya que no permiten la comprensión del lenguaje. Por tanto, uno de los objetivos de FunGramKB consiste en poder beneficiar al PLN de las investigaciones en los modelos funcionales del lenguaje, más concretamente, en la GPR a nivel léxico y en el MLC a nivel construccional. Además, como comenta Grishman (1986), la lingüística computacional parece un terreno idóneo para la verificación de las teorías lingüísticas, lo cual nos ha permitido, por ejemplo, extender el modelo teórico de la GPR con el fin de proporcionar una solución más elegante a la representación semántica léxica en un entorno multilingüe.

Por una parte, la GPR es uno de los modelos funcionales más relevantes en el panorama lingüístico actual. Este modelo gramatical adopta un enfoque comunicativo-cognitivo del lenguaje, i.e. las estructuras morfosintácticas y las reglas gramaticales deben ser explicadas en relación con sus funciones semánticas y comunicativas. En la GPR, los componentes sintácticos están directamente proyectados en términos de un algoritmo, el cual incluye un conjunto de reglas que describen la interfaz sintáctico-semántica. Como resultado, la GPR permite que el aducto textual sea representado en términos de una estructura lógica. Por ejemplo, la estructura lógica de la unidad léxica *querer* adopta la forma del estado (3).

(3) querer' (x, y)

En FunGramKB, la estructura lógica de la GPR ha sido mejorada, dando como resultado un nuevo formalismo denominado "estructura lógica conceptual" (Periñán Pascual y Mairal Usón, 2009), donde una estructura como (3) sería reemplazada por la representación (4).

(4) +WANT_00 (x_{Theme}, $y_{Referent}$)

Las principales ventajas de las estructuras lógicas conceptuales residen en dos aspectos:

(i) Son representaciones independientes de la lengua, ya que no están configuradas por unidades léxicas sino por conceptos ontológicos. Gracias a este enfoque interlingüístico, somos capaces de minimizar la redundancia de información al mismo tiempo que maximizamos la informatividad.

(ii) El motor de razonamiento posee un mayor potencial inferencial cuando las predicciones se basan en expectativas cognitivas.

Con respecto a este último punto, cuando aplicamos los mecanismos de razonamiento (cf. Periñán Pascual y Arcas Túnez 2005; Periñán Pascual y Mairal Usón 2009), es preciso que la estructura lógica conceptual sea transducida a una representación en COREL, y así enriquecer la estructura lógica conceptual con el conocimiento de los postulados de significado, los guiones, los retratos y/o las historias más relevantes para el modelo de situación descrito por el texto. Por ejemplo, la oración (5) presenta la estructura lógica conceptual (6), la cual es proyectada al esquema COREL (7).

(5) Juan quiere un cuchillo.

(6) $<_{IF}$ DEC $<_{TNS}$ PRES <+WANT_00 (%JUAN_00_{Theme}, $<_{DEF}$ + $<_{NEG}$ ∅ $<_{QNT}$ E $<_{NUM}$ SINGULAR $<_{NASP}$ COUNT (+KNIFE_$00_{Referent}$)>>>>>)>>>

(7) +(e1: pres +WANT_00 (x1: %JUAN_00)$_{Theme}$ (x2: +KNIFE_00)$_{Referent}$)

En definitiva, las estructuras lógicas conceptuales sirven como puente entre las idiosincrasias particulares codificadas en una determinada expresión lingüística y el nivel conceptual de FunGramKB.

Por otra parte, el MLC, el cual también está fundamentado en el marco de la GPR, va más allá de la gramática nuclear incorporando dimensiones del significado de larga tradición en la pragmática y el análisis del discurso. Más concretamente, el MLC reconoce cuatro niveles construccionales (i.e. argumental, implicativo, ilocutivo y discursivo) que dan forma a los cuatro Constructicones del Gramaticón. Por ejemplo, la estructura lógica conceptual (6) se genera automáticamente por medio de la Gramática Nuclear del verbo. En cambio, la oración (8) necesita además la información sobre la construcción resultativa almacenada en el Constructicón de nivel 1 con el fin de obtener la estructura lógica conceptual (9) y el esquema COREL (10).

(8) John ate the plate clean.

(9) <$_{IF}$ DEC <$_{TNS}$ PAST <do (%JOHN_00$_{Theme}$ [+EAT_00 (%JOHN _00$_{Theme}$, +PLATE_00$_{Referent}$)]) ($DIRTY_N_00$_{Attribute}$)>>>

(10) +(e1: past +EAT_00 (x1: %JOHN_00)Theme (x2: +PLATE_00)Referent (f1: (e2: +BECOME_00 (x2)Theme (x3: $DIRTY_N_00)Attribute))Result)

4 FunGramKB y el procesamiento del lenguaje natural

El procesamiento textual en el que FunGramKB esté implicado puede secuenciarse en las siguientes fases: (i) tokenización, (ii) análisis morfológico, (iii) procesamiento sintáctico-semántico, (iv) construcción de la estructura lógica conceptual, y, opcionalmente, (v) razonamiento. A continuación, describimos cada una de estas fases, cuyo diagrama de secuencia en UML se presenta en el apéndice 1. Con el fin de ilustrar todo el procesamiento, tomaremos como ejemplo la oración (5).

Durante la tokenización, el aducto se segmenta en oraciones y en palabras ortográficas, las cuales son tratadas como unidades básicas del análisis. El resultado de esta fase se presenta en (11).

(11) Juan | quiere | un | cuchillo

En el análisis morfológico, las palabras ortográficas son desprovistas de sus afijos flexivos, pero la información gramatical proporcionada por este análisis (p.ej. el tiempo, aspecto y modo del verbo) es almacenada en una matriz atributo-valor. Esta fase del procesamiento precisa la consulta no sólo de las entradas léxicas almacenadas en el Lexicón sino también de los dos componentes del Morficón: MorphoRules, donde se almacenan reglas en forma de patrones flexivos, y MorphoDB, una base de datos de formas léxicas irregulares. El resultado de esta fase se presenta en (12).

(12) Juan | querer | un | cuchillo

En el procesamiento sintáctico-semántico, el sistema resuelve los casos de desambiguación léxica y determina la estructura sintagmática del aducto, siendo ambas tareas realizadas paralelamente. Según Mahesh (1995), resultan inapropiadas para la interacción sintaxis-semántica tanto las arquitecturas secuenciales, donde un proceso de nivel inferior no puede obtener información de un proceso de nivel superior, como las arquitecturas integradas, donde los diferentes tipos de conocimiento se aplican de manera independiente. Por ello, abogamos por una

configuración paralela con un controlador interactivo que mantenga la independencia de la sintaxis y la semántica pero que también permita la comunicación bidireccional entre el analizador léxico-semántico y el analizador sintáctico. Al final de esta fase, el sistema genera una representación parentética donde los lemas de las palabras de contenido han sido reemplazados por etiquetas conceptuales, como se ilustra en (13).

(13) S(NP(n(%JUAN_00)), VP(v(+WANT_00)), (NP(det(a), n(+KNIFE_00)))

Resulta evidente que una de las tareas más complejas en esta fase es el proceso de desambiguación léxica, cuyo algoritmo puede acceder a diferentes tipos de conocimiento en FunGramKB, más concretamente aquellos almacenados en el Lexicón, el Gramaticón y la Ontología.[6]

Con el fin de construir automáticamente la estructura lógica conceptual (6) a partir de la estructura sintagmática y la información léxico-conceptual almacenada en la matriz atributo-valor, el sistema precisa ahora consultar el conocimiento almacenado en la Gramática Nuclear del Lexicón y aplicar el algoritmo de enlace de la GPR (Van Valin, 1998).

En caso de que el sistema requiera algún tipo de razonamiento con el texto, la CLS de la fase anterior es transducida automáticamente a un esquema COREL, tal y como se presentó en (7). Durante esta proyección de estructura lógica conceptual a esquema COREL, los únicos elementos de la estructura lógica conceptual que se toman en consideración son los operadores gramaticales (p.ej. pres), los conceptos de FunGramKB (p.ej. +KNIFE_00) y sus papeles temáticos (p.ej. Theme). Gracias a este esquema de COREL, y a que todas las representaciones conceptuales almacenadas en los diversos módulos cognitivos de FunGramKB están también formalizadas con este mismo lenguaje, podemos inferir más fácilmente el conocimiento relevante sobre la situación. Por ejemplo, uno de los componentes del motor de razonamiento de FunGramKB es el MicroKnowing (Microconceptual-Knowledge Spreading), el cual permite extender semánticamente el postulado de significado (cf. Periñán Pascual y Arcas Túnez 2005). Así, los conceptos en una predicación como (7) pueden expandir sus postulados de significado del tal forma que el sistema pueda inferir a través de +KNIFE_00 que la intención de Juan sea probablemente cortar algo, como se deduce del postulado de significado de +CUT_00 (14).

(14) +(e1: +SPLIT_00 (x1)Theme (x2)Referent (f1: +SCISSORS_00 ^ +KNIFE_00)Instrument)

De esta forma, nuestro modelo de comprensión del lenguaje es factible porque dos tipos de representaciones interlingüísticas están estrechamente interconectadas. Por una parte, la estructura lógica conceptual, la cual es capaz de explicar un amplio número de fenómenos lingüísticos dentro del marco de la GPR, sirve como lenguaje pivote entre el texto y su representación en COREL. Por otra parte, el esquema COREL, el cual ayuda a construir el modelo de situación del texto a partir del conocimiento conceptual en FunGramKB, sirve como lenguaje pivote entre la estructura lógica conceptual y el razonador.

Observamos que la piedra angular en un sistema del PLN en el que participe FunGramKB es la estructura lógica conceptual, ya que sirve de representación interlingüística del contenido semántico subyacente a un texto. En este contexto, utilizamos el término interlingua como se hace habitualmente en la traducción automática, i.e. una representación universal e independiente de la lengua generada a partir del texto origen y desde la cual se genera el texto destino. Por tanto, la interlingua debe ser capaz de describir directamente la realidad del mundo sin

[6] Véase Periñán Pascual y Marial Usón (2010a: 23) para una descripción del algoritmo de desambiguación léxica.

mediación de una lengua natural. Aunque nadie ha conseguido especificar las propiedades y estructuras conceptuales de una interlingua ideal, su realidad psicológica parece incuestionable, dado que un hablante es capaz de traducir entre lenguas tipológicamente diferentes.[7] Nuestra estructura lógica conceptual puede desempeñar un papel decisivo, por ejemplo, en los campos de la traducción automática y la recuperación de información.

En la traducción automática, la principal ventaja del enfoque interlingüístico radica en su economía computacional, ya que la traducción entre todos los pares de lenguas del sistema requiere sólo la traducción hacia y desde la interlingua para cada una de las lenguas. Por ejemplo, Periñán Pascual y Mairal Usón (2010a) describieron el modelo de implementación de FunGramKB como base de conocimiento de UniArab (Nolan and Salem, 2009; Salem, Hensman and Nolan, 2008a, 2008b; Salem and Nolan, 2009a, 2009b), un prototipo de traductor automático árabe-inglés. UniArab es uno de los primeros sistemas que implementa computacionalmente el modelo lingüístico de la GPR, siendo uno de sus puntos fuertes la capacidad de construir automáticamente una representación de la estructura lógica de una oración árabe. Sin embargo, uno de los principales problemas de este modelo radica precisamente en su lenguaje de representación semántica, el cual se fundamenta en la versión estándar de la estructura lógica de la GPR. Con el fin de solventar éste y otros problemas, la base de datos léxica de UniArab puede ser sustituida por FunGramKB, donde las entradas léxicas no sólo son informativamente más completas sino además sus representaciones de significado son más profundas. De esta manera, convertimos a UniArab en un auténtico sistema de traducción basado en el conocimiento.

Otro campo de aplicación de FunGramKB es la recuperación de información, cuyo objetivo consiste en seleccionar automáticamente, a partir de una colección muy extensa de documentos (i.e. textos, imágenes, audio, etc), aquéllos que sirvan como respuesta a la consulta textual del usuario. En este escenario, la estructura lógica conceptual puede desempeñar un papel fundamental como lenguaje de representación del contenido semántico del repositorio documental, orientando así el motor de búsqueda a lo que se conoce como "web semántica" (Berners-Lee 1998), donde los sistemas que procesan conocimiento requieren la creación de una semántica comprensible por la máquina que permita encontrar los documentos de la web de forma más eficiente. Por tanto, una futura aplicación de la estructura lógica conceptual consistiría en utilizarla como lenguaje de anotación semántica que posibilitara el acceso inteligente a la información de una base de datos documental, no sólo minimizando así el tiempo de búsqueda sino también mejorando la precisión y la cobertura del buscador.

5 El papel del lingüista en FunGramKB

La colaboración entre lingüistas e informáticos es necesaria si deseamos construir una aplicación de la ingeniería lingüística que vaya más allá de la resolución de pequeños problemas *ad hoc*. A este respecto, el papel del lingüista puede ser periférico o pleno dentro del equipo de investigadores del proyecto.

En caso de que el lingüista prefiera reducir su ámbito de colaboración a la base de conocimiento, en nuestro caso FunGramKB, calificaremos su integración en el proyecto del PLN como de "periférica", ya que no exigirá un intercambio continuo de conocimientos con los informáticos, dejando a éstos la gestión de la mayor parte del

[7] Entre los diversos sistemas del PLN que han utilizado un enfoque basado en la interlingua, destacamos UNITRAN (Dorr 1993), un sistema de traducción automática capaz de traducir textos en inglés, español y alemán bidireccionalmente, y MILT (Dorr, Hendler et al. 1995), un sistema tutorial para el aprendizaje de lenguas extranjeras, donde en ambos sistemas la representación interlingüística se construye a partir de una versión del modelo de "estructura conceptual léxica" propuesto por Jackendoff (1983, 1990).

proyecto. De hecho, en la mayoría de las ocasiones los lingüistas prefieren dedicarse a desarrollar las fuentes de conocimiento lingüístico (i.e. léxico y gramatical) y no lingüístico (i.e. conceptual) utilizadas por el sistema, en lugar de involucrarse en el diseño y desarrollo de todo el proyecto. Una de las razones por las cuales los propios lingüistas rehuyen a integrarse de forma plena en un proyecto del PLN suele atribuirse a la falta de conocimiento técnico. A pesar de que muchos lingüistas se sienten atraídos por las diversas aplicaciones de las tecnologías del lenguaje, el problema es que se muestran con frecuencia reacios a adquirir dicho conocimiento especializado, aunque estén acostumbrados a adquirir conocimientos pertenecientes a otras disciplinas del saber (p.ej. pedagogía, psicología, etc). No obstante, el papel del lingüista como proveedor de datos, i.e. la población de los diversos módulos que componen FunGramKB, exige del lingüista el desarrollo de destrezas como la abstracción, el razonamiento lógico y la organización y estructuración de los datos. Por tanto, a pesar de este papel periférico, FunGramKB ayuda a los lingüistas a desarrollar su capacidad de formalización del conocimiento léxico, gramatical y conceptual, sirviendo así como antesala al tratamiento computacional del lenguaje.

Por otra parte, la integración "plena" de informáticos y lingüistas en un equipo del PLN que utilice FunGramKB como base de conocimiento implica la posibilidad de que se tomen decisiones de forma compartida con respecto al diseño y desarrollo de todo el sistema del PLN. No obstante, este tipo de integración implica un cierto "cambio de perspectiva" (Llisterri 2003) que sólo se adquiere a través de una formación técnica más específica: no sólo conocimientos informáticos básicos (p.ej. lenguajes de programación, gestión de bases de datos e ingeniería del software) sino también conocimientos sobre métodos estadísticos y técnicas de aprendizaje automático utilizados en el PLN. En otras palabras, la fluidez de comunicación entre informáticos y lingüistas depende principalmente de su formación.[8] No obstante, aunque este tipo de participación implique que los lingüistas deban adquirir un cierto conocimiento especializado sobre ingeniería del software, no estamos sugiriendo en ningún momento que se impliquen en la programación del propio sistema, ya que podrían correr el riesgo de "reinventar la rueda": una cosa es conocer los fundamentos básicos de la programación informática y otra cosa muy diferente es ser un experto en algoritmia. Por tanto, la ingeniería de aplicaciones informáticas no debe ser realizada por los lingüistas, quienes deben limitarse a la producción de investigación avanzada. Como ilustra Ferrari (2004), aunque los puentes y edificios permanecen de pie gracias a algunos principios físicos, los físicos no tienen que intervenir finalmente en los proyectos de ingeniería.

6 Conclusiones

En este artículo, hemos descrito FunGramKB, una base de conocimiento léxico-conceptual diseñada para su uso en diversas aplicaciones del PLN, preferentemente en aquellas que requieran la comprensión del lenguaje y/o que precisen de un procesamiento multilingüe. Con un recurso de estas características, no sólo queremos hacer que el PLN vuelva a beneficiarse de las investigaciones en lingüística teórica (en nuestro caso, la GPR y el MLC), sino también pretendemos que esas investigaciones

[8] Igualmente, los informáticos que participen en proyectos del PLN deben poseer una sólida formación en lingüística descriptiva (Moore 2009), e incluso tener la posibilidad de estar directamente implicados en proyectos de investigación lingüística, ya que su formación computacional puede aportar una visión diferente (Wintner 2009). Por tanto, los informáticos deberían ser capaces de comprender, por ejemplo, los principales modelos en lingüística teórica y, en el caso de la comprensión del lenguaje natural, las premisas fundamentales en las investigaciones en ciencia cognitiva. Desgraciadamente, como confiesa Moore (2009), muchos informáticos se especializan en lingüística computacional sin poseer suficientes conocimientos sobre la estructuración interna de las lenguas.

lingüísticas puedan estar perfectamente integradas en el enfoque simbólico de la inteligencia artificial. Tal y como se ejemplificó, FunGramKB permite conectar la lingüística y la inteligencia artificial por medio de un puente cuyos dos pilares básicos adoptan la forma de representaciones interlingüísticas: la estructura lógica conceptual, como producto de la lingüística, y el esquema conceptual en COREL, orientado a la ingeniería del conocimiento. De esta forma, nuestro proyecto promueve la estrecha colaboración entre lingüistas e informáticos con el fin de desarrollar aplicaciones del PLN más robustas y flexibles.

Agradecimientos

Este trabajo forma parte de varios proyectos de investigación financiados por el Ministerio de Ciencia y Tecnología, códigos FFI2008-05035-C02-01, FFI2010-15983 y FFI2010-17610.

Referencias

Adarraga, Pablo (1994): "Sistemas basados en conocimiento: conceptos básicos". En: Adarraga, Pablo y Zaccagnini, José Luis (eds.) *Psicología e Inteligencia Artificial*. Madrid: Trotta, 141-186.

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

Barsalou, Lawrence W. (1991): "Deriving categories to achieve goals". En: Bower, Gordon H. (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 27. San Diego: Academic Press, 1-64.

Barsalou, Lawrence W. (2002): "Being there conceptually: simulating categories in preparation for situated action". En: Stein, Nancy L., Bauer, Patricia J. y Rabinowitz, Mitchell (eds.) *Represenation, Memory and Development: Essays in Honor of Jean Mandler*. Mahwah (NJ): Erlbaum, 1-15.

Berners-Lee, Tim (1998): "Semantic Web road map". [http://www.w3.org/DesignIssues/Semantic.html]

Dorr, Bonnie Jean (1993): *Machine Translation: A View from the Lexicon*. Cambridge (Mass.): MIT Press.

Dorr, Bonnie Jean, Hendler, Jim, Blanksteen, Scott y Migdalof, Barrie (1995): "Use of LCS and discourse for intelligent tutoring: on beyond syntax". En: Holland, Melissa, Kaplan, Jonathan y Sams, Michelle (eds.) *Intelligent Language Tutors: Balancing Theory and Technology*. Hillsdale: Lawrence Erlbaum Associates, 289-309.

Ferrari, Giacomo (2004): "State of the art in Computational Linguistics". En: Sterkenburg, Piet van (ed.) *Linguistics Today. Facing a Greater Challenge*. Amsterdam: John Benjamins, 163-186.

Fraser, Norman (1991): "Review of *Functional Grammar and the Computer* by John H. Connolly and Simon C. Dik". Computational Linguistics 17 (1), 104-106.

Gibbon, Dafydd (2000): "Computational lexicography". En: Frank van Eynde y Dafydd Gibbon (eds.) *Lexicon Development for Speech and Language Processing*. Dordrecht, Kluwer, 1-42.

Grishman, Ralph (1986): *Computational Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Halvorsen, Per-Kristian (1988): "Computer applications of linguistic theory". En: Newmeyer, Frederick (ed.) *Linguistics: The Cambridge Survey. II Linguistic Theory: Extensions and Implications*. Cambridge: Cambridge University Press, 198-219.

Jackendoff, Ray (1983): *Semantics and Cognition*. Cambridge (Mass.): MIT Press.

Jackendoff, Ray (1990): *Semantic Structures.* Cambridge (Mass.): MIT Press.

Lenci, Alessandro *et al.* (2000): "SIMPLE: A general framework for the development of multilingual lexicons". International Journal of Lexicography 13 (4), 249-263.

Llisterri Boix, Joaquim (2003): "Lingüística y tecnologías del lenguaje". En *Lynx. Panorámica de Estudios Lingüísticos* 2, 9-71.

Mahesh, Kavi (1995): *Syntax-semantics Interaction in Sentence Understanding*, tesis doctoral, Georgia Institute of Technology, Atlanta.

Mairal Usón, Ricardo y Ruiz de Mendoza, Francisco José (2009): "Levels of description and explanation in meaning construction". En: Butler, Christopher y Martín, Javier (eds.) *Deconstructing Constructions*. Amsterdam: John Benjamins, 153-198.

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

Moore, Robert C. (2009): "What do computational linguists need to know about linguistics?". En: *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*, Athens, Association for Computational Linguistics, 41-42.

Nolan, Brian y Salem, Yasser (2009): UniArab: an RRG Arabic-to-English machine translation software. En: *Proceedings of the Role and reference Grammar International Conference*, Berkeley.

Periñán Pascual, Carlos y Arcas Túnez, Francisco (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, 38-42.

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

Periñán Pascual, Carlos y Arcas Túnez, Francisco (2007): "Cognitive modules of an NLP knowledge base for language understanding". *Procesamiento del Lenguaje Natural* 39, 197-204.

Periñán Pascual, Carlos y Arcas Túnez, Francisco (2008): "A cognitive approach to qualities for NLP". *Procesamiento del Lenguaje Natural* 41, 137-144.

Periñán Pascual, Carlos y Arcas Túnez, Francisco (2010a): "Ontological commitments in FunGramKB". *Procesamiento del Lenguaje Natural* 44, 27-34.

Periñán Pascual, Carlos y Arcas Túnez, Francisco (2010b): "The Architecture of FunGramKB". En: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Malta, ELRA, 2667-2674.

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

Periñán Pascual, Carlos y Mairal Usón, Ricardo (2010a): "Enhancing UniArab with FunGramKB". *Procesamiento del Lenguaje Natural* 44, 19-26.

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

Ruiz de Mendoza Ibáñez, Francisco José y Mairal Usón, Ricardo (2008): "Levels of description and constraining factors in meaning construction: an introduction to the Lexical Constructional Model". *Folia Linguistica* 42 (2), 355–400.

Salem, Yasser, Hensman, Arnold y Nolan, Brian (2008a): "Implementing Arabic-to-English machine translation using the Role and Reference Grammar linguistic model". En: *Proceedings of the 8th Annual International Conference on Information Technology and Telecommunication,* Galway, Irlanda.

Salem, Yasser, Hensman, Arnold y Nolan, Brian (2008b): "Towards Arabic to English machine translation". *ITB Journal* 17, 20-31.

Salem, Yasser y Nolan, Brian (2009a): "Designing an XML lexicon architecture for Arabic machine translation based on Role and Reference Grammar". En:

Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egipto.

Salem, Yasser y Nolan, Brian (2009b): "UniArab: a universal machine translator system for Arabic Based on Role and Reference Grammar". En: *Proceedings of the 31st Annual Meeting of the Linguistics Association of Germany*.

Van Valin, Robert D. Jr. (2005): *Exploring the Syntax-Semantic Interface*. Cambridge: Cambridge University Press.

Van Valin, Robert D. Jr. y LaPolla, Randy (1997): *Syntax: Structure, Meaning, and Function*. Cambridge: Cambridge University Press.

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

Wintner, Shuly (2009): "What science underlies natural language engineering?". *Computational Linguistics* 35 (4), 641-644.

Yeh, Wenchi y Barsalou, Lawrence W. (2006): "The situated nature of concepts". *American Journal of Psychology* 119 (3), 349-384.

Apéndice 1. Diagrama de secuencia en el procesamiento textual con FunGramKB.