FIRST DRAFT

La dimensión computacional de la GPR: la estructura lógica conceptual y su aplicación en el procesamiento del lenguaje natural

Carlos Periñán Pascual
Universidad Católica San Antonio
Murcia, España
jcperinan@pdi.ucam.edu

Ricardo Mairal Usón
UNED
Madrid, España
rmairal@flog.uned.es

1- Introducción

Como continuación al capítulo XXX de este volumen donde hemos presentado las ELC como alternativa a las estructuras lógicas de la GPR, es nuestro objetivo en este trabajo detenernos en el alcance explicativo que estas representaciones tienen en el ámbito del procesamiento del lenguaje natural, lo cual nos permite además plantear la adecuación computacional de la GPR, un aspecto que, sorprendemente, no ha recibido mucha atención.

Argumentábamos a favor de un enfoque conceptual, para lo cual propusimos la inclusión de una base de conocimiento léxico-conceptual, i.e. FunGramKB, en lugar de una simple base de datos léxica. Dedicamos la Sección 2 a describir la arquitectura de FunGramKB, con especial atención al módulo ontológico y al módulo léxico. Para comprender cómo se construye una ELC, es fundamental tener una visión aproximada de estos dos componentes así como de la arquitectura general de la base de conocimiento. En la Sección 3, retomamos una cuestión que informalmente avanzamos en el capítulo XXX: el papel de la ELC como interlingua que sirve como lenguaje pivote entre el aducto y la representación conceptual en COREL, sobre la cual se aplicará el motor de razonamiento. Finalmente, la Sección 4 explora el alcance explicativo de la ELC en el ámbito del procesamiento del lenguaje natural y, más en particular, en la traducción automática y en la recuperación de la información.

2- FunGramKB

2.1 Definición y estructuración modular

FunGramKB[1] es una base de conocimiento léxico-conceptual multipropósito diseñada principalmente para su uso en sistemas del procesamiento del lenguaje natural (PLN), y más concretamente, para aplicaciones que requieran la comprensión de la lengua. 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. la recuperación y la extracción de información, la traducción automática, los 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:

- El Lexicón almacena información morfosintáctica, pragmática y colocacional sobre las unidades léxicas.
- El Morficón asiste al analizador y al generador en el tratamiento de los casos de morfología flexiva.

Nivel gramatical:

- El Gramaticón almacena los esquemas construccionales que pueden ser utilizados por el algoritmo de enlace sintáctico-semántico de la GPR (Van Valin and LaPolla, 1997; Van Valin, 2005).

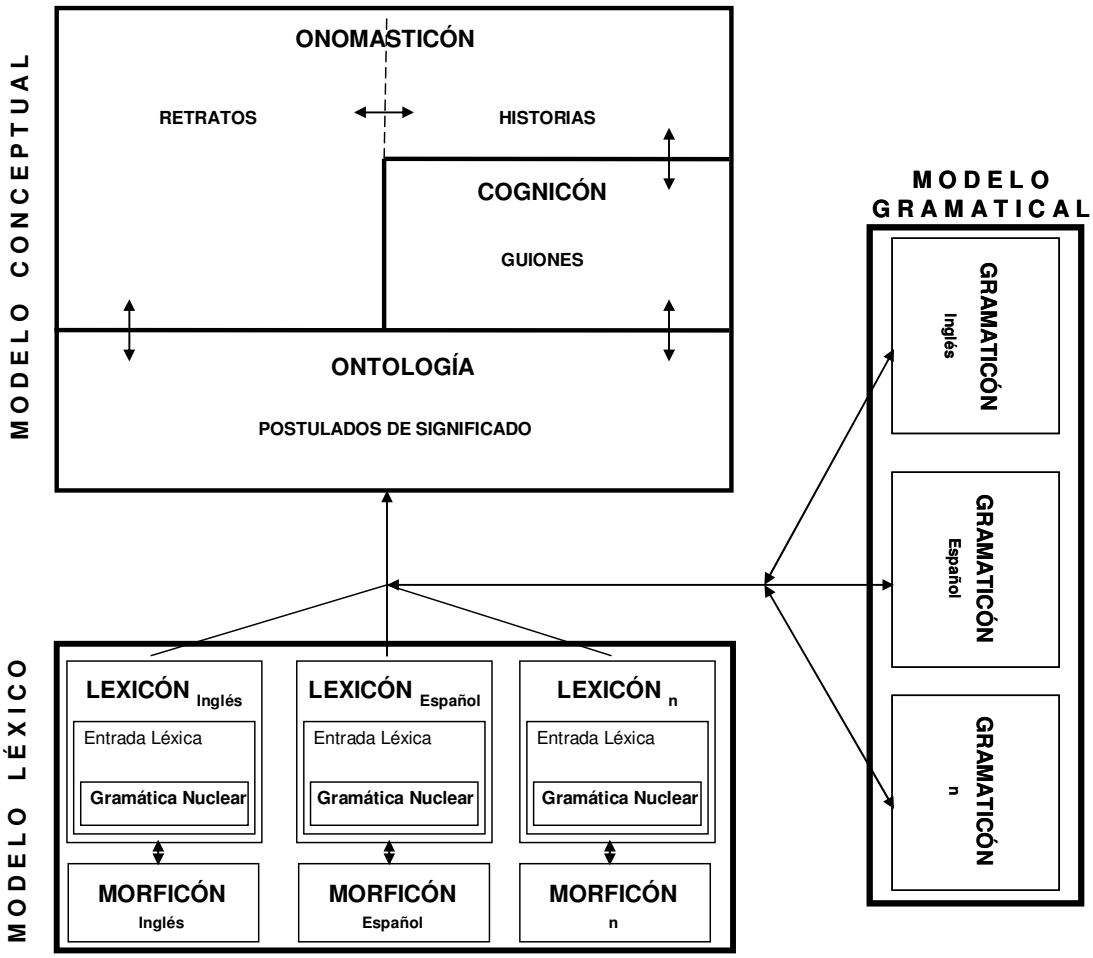
[1] www.fungramkb.com
[2] Actualmente, FunGramKB ha sido modelada para poder trabajar con siete lenguas: alemán, búlgaro, catalán, español, francés, inglés e italiano.
[3] Para una información más detallada sobre el conocimiento almacenado en FunGramKB, léanse Mairal Usón y Periñán Pascual (2009) con respecto al nivel léxico, y Periñán Pascual y Arcas Túnez (2004, 2007a, 2008, 2010a, 2010b) y Periñán Pascual y Mairal Usón (2010b) a cerca del nivel conceptual.

Nivel conceptual:

- 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 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 lógica de intervalos de Allen (1983).
- El Onomasticón almacena el conocimiento enciclopédico sobre instancias de entidades y eventos, tales como Chicago o La Segunda Guerra Mundial. Este módulo almacena su conocimiento por medio de dos tipos diferentes de esquemas (i.e. retratos e historias), ya que las instancias pueden ser descritas sincrónica o diacrónicamente.

En la arquitectura de FunGramKB (Figura 1), 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. 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. En los siguientes dos subapartados, describimos los modelos ontológico y léxico con el fin de presentar el escenario cuyo conocimiento ayuda a generar automáticamente las estructuras lógicas conceptuales (cf. capítulo xxx de este volumen).

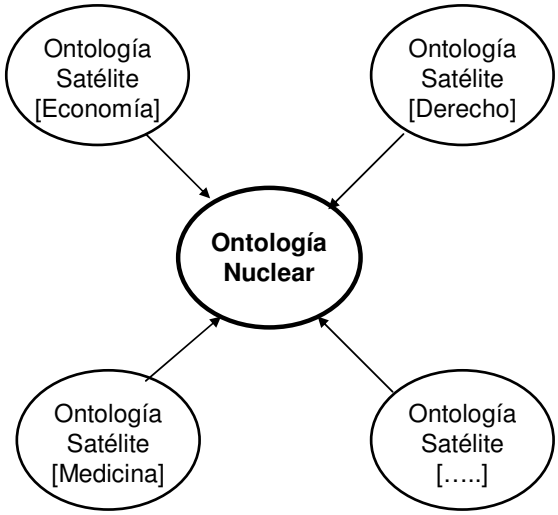
Figura 1. La arquitectura modular de FunGramKB.



2.2 El modelo ontológico

El modelo ontológico consiste en dos componentes: un módulo de propósito general (i.e. la Ontología Nuclear) y varios módulos terminológicos específicos del dominio (i.e. las Ontologías Satélites). El resto de este subapartado lo dedicamos exclusivamente a la descripción de la Ontología Nuclear, por tratarse ésta del pivote sobre el cual giran las Ontologías Satélites (Figura 2).

Figura 2. La Ontología Nuclear y las Ontologías Satélites.



La Ontología Nuclear de FunGramKB, la cual se concibe como una taxonomía conceptual IS-A que permite la herencia múltiple no monotónica,[4] distingue tres niveles conceptuales diferentes, cada uno de los cuales contiene conceptos de diferente naturaleza:

(i) Los *metaconceptos* representan dimensiones cognitivas dentro de las cuales se organizan el resto de unidades ontológicas. Los metaconceptos son el resultado del análisis del nivel superior de las principales ontologías, por ej. SUMO, DOLCE, GUM, Mikrokosmos, SIMPLE etc. Algunos ejemplos de metaconceptos son #ABSTRACT, #MOTION y #TEMPORAL;

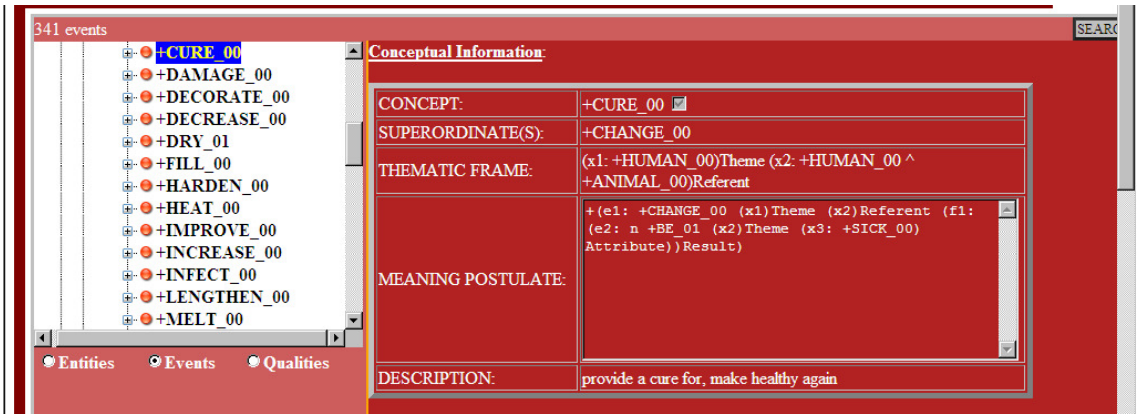
(ii) Los *conceptos básicos* constituyen el nivel intermedio de la Ontología, los cuales se obtuvieron a partir del vocabulario definitorio del *Longman Dictionary of Contemporary English* (Procter, 1978). El inventario final de conceptos básicos no se elaboró como resultado directo de una proyección léxico-conceptual, sino más bien como el producto de un complejo proceso de cuatro fases: conceptualización, jerarquización, remodelación y refinamiento. Algunos ejemplos de conceptos básicos son +FEEL_00, +PERCEIVE_00 y +WANT_00;

[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.

(iii) Los *conceptos terminales* representan los nodos finales de la estructuración jerárquica conceptual. Algunos ejemplos de conceptos terminales son $ADAPT_00, $FLUCTUATE_00 y $MEDITATE_00.

Los conceptos básicos y terminales no se conciben como unidades atómicas, sino poseen una serie de propiedades semánticas, entre las que destacamos los marcos temáticos y los postulados de significado (Figura 3), los cuales desempeñan un papel crucial en el vínculo léxico-conceptual (Periñán Pascual y Mairal Usón, 2009).

Figura 3. La entrada conceptual en la Ontología Nuclear.



Por una parte, cada evento en la Ontología tiene asignado un único marco temático, i.e. un constructo conceptual que indica el número y tipo de participantes implicados en la situación cognitiva prototípica descrita por el concepto. A modo de ilustración, presentamos el marco temático de +CURE_00, a la que se vinculan unidades léxicas como *cure, heal* [inglés], *curar, sanar* [español], *curare, sanare, guarire* [italiano], *behandeln, heilen, kurieren* [alemán] o *guérir* [francés]:

(1) (x1: +HUMAN_00)Theme (x2: +HUMAN_00 ^ +ANIMAL_00)Referent

Los participantes de los marcos temáticos pueden incluir aquellas preferencias de selección que se encuentren típicamente implicadas en la situación cognitiva. De esta forma, el marco temático (1) describe una situación en la que típicamente están involucrados dos participantes: una persona (x_1) cura a otra persona o animal (x_2).

Por otra parte, los conceptos también están provistos de un postulado de significado, el cual se representa como un conjunto de una o más predicaciones (e_1, e_2...

e_n) conectadas de forma lógica, i.e. constructos conceptuales que portan los rasgos genéricos de los conceptos.[5] En el caso del concepto +CURE_00, su postulado de significado se representa como (2), siendo (3) el equivalente de traducción al español:

(2) +(e1: +CHANGE_00 (x1)Theme (x2)Referent (f1: (e2: n +BE_01 (x2)Theme (x3: +SICK_00)Attribute))Result)

(3) Una persona cambia el estado de otra persona o animal, haciendo que esta última entidad deje de estar enferma.

Si la carga semántica de este concepto, y por consiguiente de todas sus correspondientes unidades léxicas, recayese únicamente en el marco temático, entonces no estaríamos describiendo realmente el contenido conceptual de estas unidades léxicas. A diferencia de algunos otros modelos ontológicos aplicados al PLN—por ejemplo, EuroWordNet (Vossen, 1998) o SIMPLE (Lenci y otros, 2000), FunGramKB adopta un enfoque orientado a la semántica profunda, donde se enfatiza nuestro compromiso de proporcionar definiciones por medio de postulados de significado.[6]

Como explicaremos posteriormente, mientras el marco temático desempeña un papel crucial en la construcción automática de la estructura lógica conceptual,[7] los postulados de significado sirven para enriquecer sustancialmente la carga semántica del esquema conceptual originado a partir de la estructura lógica conceptual.[8]

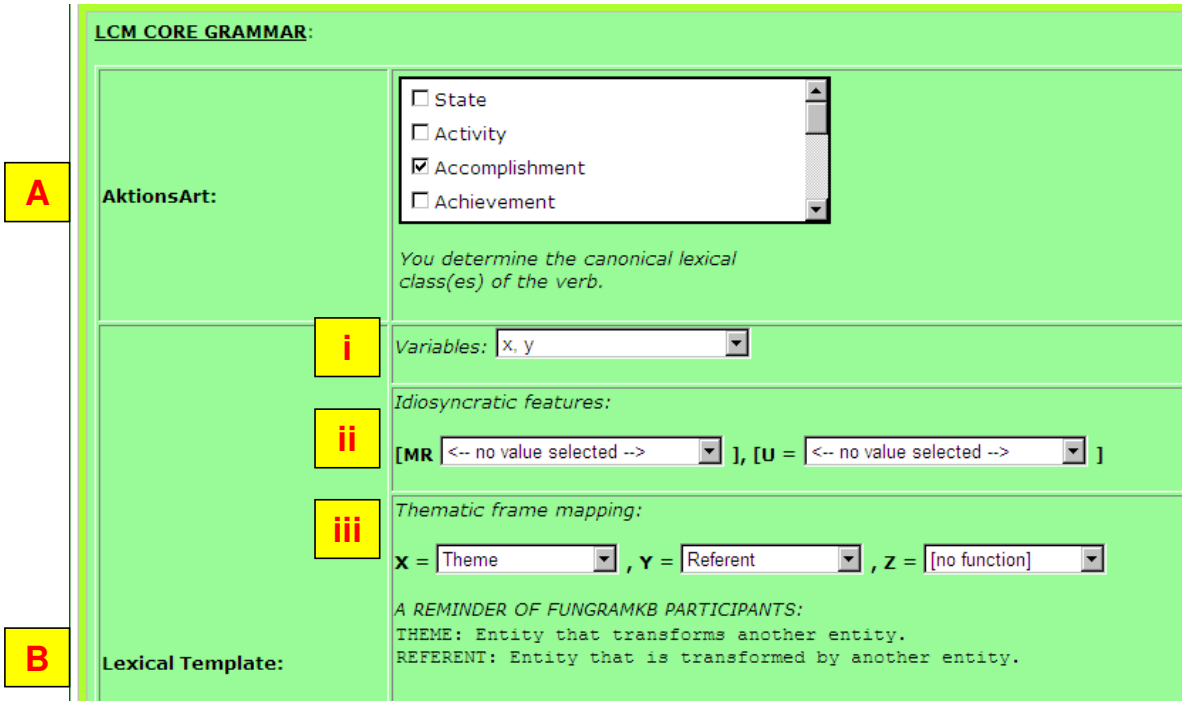
2.3 El modelo léxico

El Lexicón de FunGramKB ha sido especialmente diseñado para ser explotado por la GPR. Más concretamente, la Gramática Nuclear de la entrada léxica de los verbos contiene aquellos atributos cuyos valores permiten al sistema construir automáticamente la estructura lógica conceptual (Figura 4).

---

[5] Periñán Pascual y Arcas Túnez (2004) describen la gramática formal de las predicaciones válidas para los postulados de significado en FunGramKB.
[6] Periñán Pascual y Arcas Túnez (2007b) presentan las ventajas del enfoque de la semántica profunda frente al modelo relacional del significado tomando como ejemplos postulados de significado de FunGramKB.
[7] Véase el apartado 2.3.
[8] Véase el apartado 3.

Figure 4. La Gramática Nuclear en el Lexicón.



A continuación, describimos los principales atributos que configuran esta Gramática Nuclear:

A- *Aktionsart*

Cada unidad léxica tiene asignada uno o más *Aktionsarten* a partir del inventario de clases verbales de la GPR.

B- Plantilla léxica

(i) Variables. Una o más variables (i.e. *x, y, z*) representan los argumentos prototípicos de la unidad léxica. El número de variables viene determinado por aquel *Aktionsart* seleccionado con el mayor número de argumentos. Siguiendo con el enfoque de la GPR, las entradas léxicas no incluyen rasgos de subcategorización para los argumentos (p.ej. realización sintagmática o función sintáctica), sino simplemente el número de argumentos.

(ii) Proyección de variables. Cada variable en la entrada léxica de un verbo se liga a uno y sólo uno de los participantes en el marco temático del concepto al que está vinculada la unidad léxica. Esto no implica que todos los participantes del marco temático deban estar ligados a las variables de la plantilla léxica. Por ejemplo, *golpear* tiene asignada dos variables en su

estructura lógica, pero el marco temático del concepto +HIT_00 tiene cuatro argumentos.

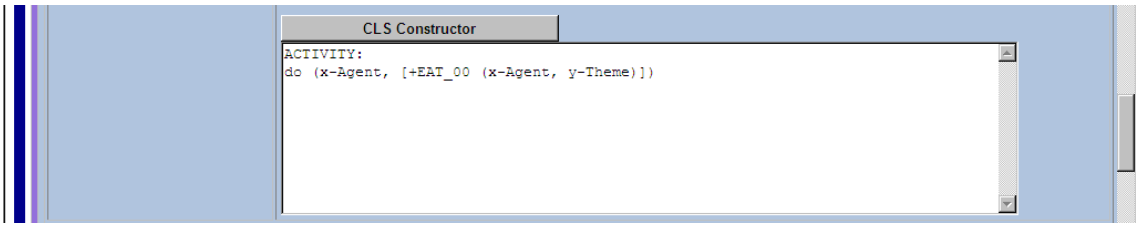
(x1: +HUMAN_00 ^ +ANIMAL_00)Agent (x2: +BODY_PART_00 ^ +STICK_00 ^ +VEHICLE_00)Theme (x3)Origin (x4)Goal

Los únicos papeles temáticos que se asignan a las variables de la estructura lógica conceptual son *Agent* y *Goal*. Esto demuestra que, durante el proceso de construcción de los marcos temáticos, los ingenieros del conocimiento no se preocupan en absoluto de los fenómenos lingüísticos, sino más bien toman sus decisiones teniendo en cuenta exclusivamente la situación cognitiva descrita por el concepto.

(iii) Rasgos idiosincrásicos (opcional). Estos rasgos, los cuales se expresan por medio del atributo [MR=α] donde α puede ser 0, 1 ó 2, sirven para indicar que una unidad léxica no sigue los principios defectivos para la asignación de los macropapeles.

A partir de esta información en la Gramática Nuclear del verbo, FunGramKB es capaz de generar automáticamente su estructura lógica conceptual básica (Figura 5).

Figura 5. La estructura lógica conceptual de *comer*.



3- La estructura lógica conceptual y el motor de razonamiento

En el capítulo xxxx, hemos podido comprobar la relevancia de la estructura lógica conceptual como esquema que sustituye a la estructura lógica convencional dentro del marco de la GPR. En los siguientes dos apartados, destacamos la importancia de la estructura lógica conceptual en el campo del PLN.

El desarrollo de sistemas de comprensión del lenguaje natural suele

obstaculizarse por dos problemas principales. En primer lugar, se precisa una base de conocimiento extensa, la cual no sólo contenga el conocimiento semántico de las palabras sino también el conocimiento del mundo. En segundo lugar, hemos de implementar un razonador automático que pueda obtener el conocimiento implícito en un texto de entrada. Con respecto al razonamiento automatizado en FunGramKB, actualmente se está desarrollando un motor que será capaz de inferir conclusiones a partir de la información sobre los hechos del mundo real y el conocimiento almacenado en los módulos conceptuales. Este razonador se implementará computacionalmente por medio de dos módulos: el MicroKnowing y el Constructor de Presuposiciones. El MicroKnowing (*Microconceptual-Knowledge Spreading*) es un prerazonador multinivel para la construcción del postulado de significado extendido de un concepto, lo cual nos ayuda a inferir el conocimiento implícito en un texto.[9] El repositorio de salida del MicroKnowing actúa como un espacio conceptual donde el Constructor de Presuposiciones identifica las predicaciones más salientes para explicar la interacción contextual de la información. La salida del Constructor de Presuposiciones configura a su vez la memoria de trabajo de la aplicación, a través de la cual el sistema es capaz de reconstruir el modelo de situación referencial del texto de entrada.

No obstante, nuestro modelo de comprensión del lenguaje natural sólo es factible siempre y cuando se interrelacionen dos tipos de representaciones interlingüísticas:

(i) la estructura lógica conceptual, la cual sirve como lenguaje pivote entre el texto de entrada y su correspondiente representación en COREL, y
(ii) el esquema conceptual en COREL, el cual sirve como lenguaje pivote entre la estructura lógica conceptual y el razonador automático.

Con el fin de ilustrar esta interrelación entre la estructura lógica conceptual y los esquemas conceptuales de COREL, supongamos que el texto de entrada adopta la forma de la oración (15), cuya estructura lógica conceptual es (16):

[9] El MicroKnowing tiene lugar en un escenario multinivel, ya que ejecuta iterativamente dos tipos de mecanismos de razonamiento: la herencia y la inferencia. Entendemos la herencia como el fenómeno que implica la transferencia de una o más predicaciones desde un concepto ontológico superordinado a uno subordinado; en cambio, nuestro mecanismo de inferencia se basa en los constructos compartidos entre predicaciones vinculadas a unidades conceptuales que no participan en la misma relación de subsunción dentro de la ontología. Periñán Pascual y Arcas Túnez (2005) proporcionan una descripción precisa del funcionamiento del MicroKnowing en FunGramKB.

(4) María golpeó a Juan.

(5) $<_{IF}$ DECL $<_{TNS}$ PAST < do (%MARIA_00_{Agent}, [+HIT_00 (%MARIA_00_{Agent}, %JUAN_00_{Theme})]>>>

Con el propósito de aplicar la tarea de razonamiento sobre el texto de entrada, es preciso que la estructura lógica conceptual se "traduzca" automáticamente a un esquema conceptual en COREL, para que de esta forma la estructura lógica conceptual se enriquezca del conocimiento de los postulados de significado, los guiones, los retratos y las historias almacenados en FunGramKB. A lo largo de este proceso de proyección a COREL, los únicos elementos relevantes de la estructura lógica conceptual son los operadores gramaticales, los conceptos de FunGramKB y los papeles temáticos. Así, la estructura lógica conceptual (5) da como resultado la predicación (6):

(6) +(e1: past +HIT_00 (x1: %MARIA_00)Agent (x2)Theme (x3)Origin (x4: %JUAN_00)Goal)

Una vez reconstruido el esquema conceptual correspondiente, el siguiente paso consiste en obtener el conocimiento implícito que nos permita interpretar el texto de entrada. Ahora bien, el razonamiento práctico para aplicaciones informáticas reales no puede basarse exclusivamente en el conocimiento semántico, sino que requiere además el conocimiento del mundo, el conocimiento situacional, etc (Bos, 2005). Por ejemplo, tras aplicar el MicroKnowing, el esquema conceptual (6) incorpora las predicaciones (7), (8) y (9) a través del mecanismo de la herencia, es decir, a partir del postulado de significado del propio concepto +HIT_00 y de los de sus conceptos superordinados (i.e. +PUT_00 y +MOVE_00) hasta alcanzar su concepto básico raíz (i.e. +DO_00).

(7) +(e1: +PUT_00 (x1: %MARIA_00)Agent (x2: +CORPUSCULAR_00)Theme (x3)Origin (x4: %JUAN_00)Goal (f1: +FAST_00)Speed (f2: +HARD_00)Manner)

(8) +(e1: +MOVE_00 (x1: %MARIA_00)Agent (x2: +CORPUSCULAR_00)Theme (x5)Location (x3)Origin (x4: %JUAN_00)Goal (f1: (e2: +BE_02 (x2)Theme (x4)Location (f2: +ON_00 ^

+IN_00)Position))Result)

(9) +(e1: +DO_00 (x1: %MARIA_00)Theme (x6)Referent (f1: x3)Location (f2: (e2: +BE_02 (x2: +CORPUSCULAR_00)Theme (x4)Location))Condition (f3: (e3: n +BE_02 (x2)Theme (x4)Location))Result (f4: (e4: +BE_02 (x2)Theme (x5: %JUAN_00)Location))Purpose)

En realidad, estos postulados de significado son el producto de integrar las preferencias de selección de sus correspondientes marcos temáticos, aunque sustituyendo en ocasiones los valores por defecto por aquellos más específicos provenientes del propio texto de entrada (p.ej. %MARIA_00 y %JUAN_00) o de otros postulados de significado (p.ej. el concepto +CORPUSCULAR_00 en las últimas dos predicaciones viene inferido de la primera).[10] Gracias a esta expansión conceptual de la predicación (6), el sistema sabe ahora que prototípicamente:

(i) María hizo que un objeto tridimensional fuera desde un punto de origen hasta un punto de destino que resultó ser Juan (predicación 9),
(ii) entrando en contacto el objeto tridimensional con Juan (predicación 8), y
(iii) y realizando esta acción de forma rápida y fuerte (predicación 7).

4- La estructura lógica conceptual y el procesamiento del lenguaje natural

4.1 La traducción automática

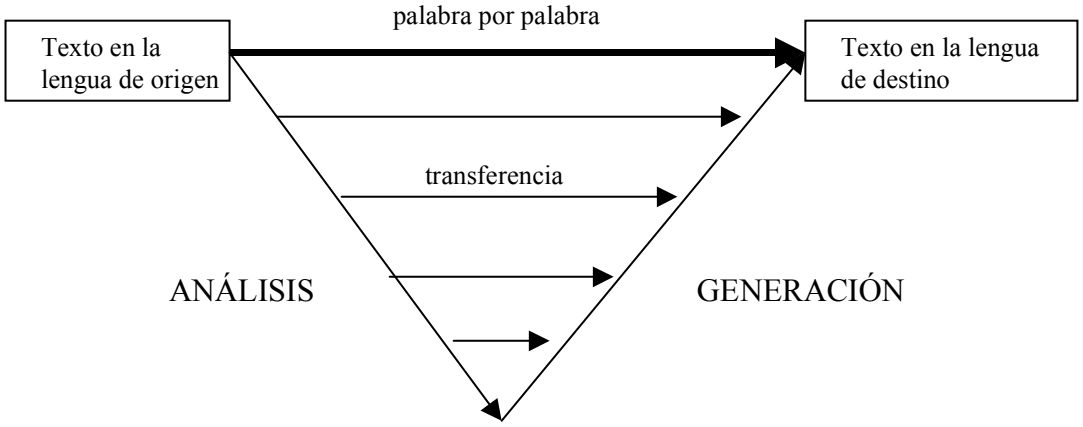
Uno de los primeros intentos de implementar la estructura lógica conceptual en un sistema del PLN ha consistido en proponer FunGramKB como base de conocimiento de UniArab (Nolan and Salem, 2009; Salem, Hensman and Nolan, 2008a, 2008b; Salem and Nolan, 2009a, 2009b), un prototipo de traductor automático árabe-inglés. Con el fin de comprender mejor el alcance de nuestra estructura lógica conceptual, es preciso describir brevemente los principales modelos de traducción automática sobre los que podríamos trabajar.

[10] Esta proyección de preferencias de selección entre predicaciones de postulados de significado diferentes resulta muy fácil porque los conceptos implicados poseen esquemas temáticos muy parecidos: Agent + Theme + Origin + Goal para +HIT_00 y +PUT_00, y Agent + Theme + Location + Origin + Goal para +MOVE_00.

Los sistemas de traducción basados en reglas[11] pueden adoptar uno de los siguientes tres enfoques: directo, transformativo e interlingüístico. De los deficientes resultados obtenidos por el enfoque directo, donde cada una de las palabras del texto de entrada se traduce "directamente" a su equivalente de traducción para posteriormente aplicar un reajuste gramatical, pronto se pasó a los sistemas de traducción basados en la transferencia. En este enfoque transformativo, tiene lugar una "transferencia" en el nivel léxico-estructural que hace corresponder las estructuras del texto fuente con estructuras en la lengua destino. La mayor parte del procesamiento que ocurre en este tipo de traductores se apoya en la información comparativa sobre pares específicos de lenguas. Una alternativa al enfoque transformativo consiste en "traducir" el texto fuente a una interlingua, la cual sea capaz de representar el "significado" de ese texto. En el enfoque interlingüístico, no hay reglas de transferencia, sino que las rutinas de traducción sintáctica y léxico-semántica proyectan sistemáticamente la estructura superficial del texto de entrada a la interlingua, y de aquí a la estructura superficial del texto de salida. Como muestra la conocida pirámide de Vauquois (1968) (Figura 8), una interlingua permite que el tamaño del módulo de transferencia se reduzca a cero. Por tanto, este modelo de traducción automática presenta sólo dos fases: el análisis del texto fuente y la generación del texto destino.

[11] Existe otro tipo de sistemas de traducción de automática que, ajenos a la utilización de reglas, se caracterizan por procesar grandes cantidades de datos en forma de córpora paralelos. Entre estos enfoques empíricos, destacan principalmente dos modelos: los modelos estadísticos, basados en el cálculo de probabilidades en la ocurrencia de las palabras, y los modelos basados en ejemplos, los cuales se apoyan en la reutilización de traducciones ya existentes como base para las nuevas traducciones.

Figura 8. Los enfoques directo, transformativo e interlingüístico 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. En cambio, el principal argumento que suele esgrimirse en contra de esa primera generación de modelo interlingüístico es la dificultad de diseñar una única interlingua que pueda ser semánticamente útil para todas las lenguas. No obstante, en la década de los 90, resurgió el modelo interlingüístico con una segunda generación de sistemas que constituyen lo que se conoce como "traducción automática basada en el conocimiento", donde el lenguaje de representación del significado de un texto se describe por medio de una ontología. La conexión entre la representación semántica y la ontología es proporcionada por el lexicón, donde los significados de las palabras se definen a través de sus proyecciones a las unidades ontológicas. Esta ontología independiente de la lengua no sólo proporciona las unidades conceptuales que constituyen la representación semántica sino también define un conjunto de operaciones de composición que permiten combinar esas unidades conceptuales con el fin de representar significados más complejos.

UniArab irrumpe en este panorama de la traducción automática con un modelo interlingüístico fundamentado en la GPR.[12] Precisamente, uno de los puntos fuertes de UniArab radica en su capacidad de construir automáticamente una representación

[12] De hecho, UniArab es uno de los primeros sistemas que implementa computacionalmente el modelo lingüístico de la GPR.

precisa de la estructura lógica de una oración árabe. Tomando el ejemplo de Salem y Nolan (2009a), la estructura lógica (10) se construye a partir de la oración (11):[13]

(10) <TNS:PAST[do'(Khalid,[read'(Khalid,(book)])]>

(11) قرأ خالد الكتاب

Actualmente, UniArab cubre una selección suficientemente representativa de palabras, aunque utilizadas en estructuras oracionales intransitivas, transitivas y ditransitivas sintácticamente poco complejas. No obstante, como afirman Salem y Nolan (2009b), los resultados de evaluación de UniArab han resultado bastante prometedores, proporcionando incluso traducciones más precisas y gramaticalmente más correctas que las obtenidas a través de traductores como Google (2009) and Microsoft (2009).

A pesar del incipiente éxito de UniArab, 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. El inadecuado tratamiento de la semántica léxica en UniArab proviene en parte del propio modelo funcional sobre el que se fundamenta. Por ejemplo, UniArab evita el problema de la desambiguación semántica adoptando el ingenuo enfoque de que cada palabra tiene asignada un único significado. Con el fin de solventar este 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. Al final del procesamiento sintáctico-semántico, la nueva versión de UniArab puede generar una representación sintáctica del texto de entrada donde los lemas hayan sido reemplazados por unidades conceptuales de la Ontología de FunGramKB. En el caso de (11), la salida sería la representación parentética (12), la cual generaría a su vez la estructura lógica conceptual (13).[14]

[13] En español, *Khalid leyó el libro*.
[14] Periñán Pascual y Mairal Usón (2010a) describen las diversas fases de procesamiento que intervienen en el nuevo modelo de UniArab.

(12) S(NP(n(%KHALID_00)), VP(v(+READ_00), NP(det(the), n(+BOOK_00))))

(13) $<_{IF}$ DECL $<_{TNS}$ PAST < do ($KHALID_00_{Theme}$, [+READ_00 ($KHALID_00_{Theme}$, $+BOOK_00_{Referent}$)] & INGR +READ_00 ($+BOOK_00_{Referent}$)>>>

Este cambio de la estructura lógica inspirada en el modelo clásico de la GPR (i.e. modelo interlingüístico de primera generación) por el enfoque conceptualista de la estructura lógica conceptual (i.e. modelo interlingüístico de segunda generación) permite que UniArab pueda ser fácilmente expandido para afrontar textos multilingües gramaticalmente más complejos, dando como resultado un sistema que se ha denominado UniLing (Nolan, Mairal Usón y Periñán Pascual, 2009).

4.2 La recuperación de información

Otro campo de aplicación de FunGramKB en el que la estructura lógica conceptual puede desempeñar un papel decisivo 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[15].

En la actualidad, la base de datos documental por excelencia es Internet y los sistemas de recuperación de información adoptan la forma de buscadores como *Google, Yahoo, Altavista*, etc. El primer problema de este tipo de sistemas de recuperación de información tiene lugar cuando el resultado de la búsqueda es una lista de documentos excesivamente larga, o bien no contiene ningún documento relevante de acuerdo con la consulta del usuario. Ambos casos son lamentablemente bastante frecuentes, debido a que los usuarios formulan sus consultas de forma poco precisa, además del hecho de que carecemos de métodos de almacenamiento y codificación que permitan convertir este acopio de datos en información organizada y estructurada (i.e. conocimiento). Un segundo problema asociado a los buscadores de Internet es la dimensión multilingüe en la que operan. Mientras el inglés es todavía la lengua predominante en Internet, el número de páginas *web* escritas en otros idiomas es cada vez mayor. La necesidad de

[15] A pesar de la popularidad de la etiqueta "recuperación de información", un término más idóneo sería "recuperación documental" (Harman, Schäuble *et al.* 1995).

formular consultas en una lengua diferente a la utilizada en la redacción de los documentos es cada vez más acuciante. Ante estos dos problemas, el reto actual de las tecnologías lingüísticas suelen centrarse en hacer que Internet pueda funcionar realmente como una gran base de conocimiento a la cual podamos acceder de manera eficaz y sin limitaciones idiomáticas.

Antes de presentar nuestro enfoque de recuperación de información basado en la estructura lógica conceptual, es preciso describir las diversas fases que suelen distinguirse a lo largo del procesamiento realizado por estos sistemas (cf. Gonzalo Arroyo y Verdejo Maíllo 2003; Tzoukermann, Klavans *et al.* 2003):

(i) el procesamiento de los documentos

En primer lugar, se identifican los índices, o palabras claves, que describen el contenido del texto[16] y se representan estos índices de forma computacionalmente eficiente. Una vez construida una lista preliminar de índices, se asigna un peso a cada uno de ellos, donde los valores más altos identifican las palabras claves más relevantes.[17]

(ii) el procesamiento de la consulta

En segundo lugar, el sistema relaciona el inventario de índices con la representación de la consulta, para lo cual se utilizan generalmente técnicas probabilísticas. Actualmente la estrategia más frecuente para representar la consulta del usuario es una combinación de palabras clave y operadores booleanos.

(iii) la presentación de los resultados

Finalmente, el resultado de la búsqueda se presenta como una lista de documentos ordenada decrecientemente de acuerdo con su relevancia.

[16] La indización de textos puede realizarse manualmente a partir de índices predefinidos, o automáticamente a través de técnicas propias de los sistemas de extracción de información. Uno de los aspectos más complejos del proceso de indización automática es encontrar la correspondencia entre las unidades léxicas y los índices, presentándose casos problemáticos como las variantes flexivas y derivativas, las expresiones multipalabra o los sinónimos de una unidad léxica.

[17] La asignación de los pesos se realiza de acuerdo a una serie de criterios de relevancia. Por ejemplo, si el índice aparece en todo los documentos de la base de datos, este índice no servirá para discriminar un documento de otro. En cambio, un índice es mucho más significativo si sólo aparece en unos pocos documentos de la colección.

En aquellos entornos de búsqueda de información en los que la lengua en que está formulada la consulta no coincide con la lengua en la que están redactados los documentos, como suele ocurrir en Internet, hablamos de "recuperación de información multilingüe". Este tipo de sistemas puede adoptar uno de los siguientes enfoques (Gonzalo Arroyo y Verdejo Maíllo 2003): traducción de los documentos, traducción de las consultas o uso de una interlingua. En este sentido, la mayoría de las investigaciones se han centrado en la traducción de la consulta a cada una de las lenguas utilizadas en la base documental. Sin embargo, teóricamente mejor es el enfoque que permite representar tanto los índices documentales como la consulta del usuario de forma interlingüística, i.e. sin mediación de una lengua natural. A pesar de las ventajas evidentes de este enfoque, todavía se está investigando cómo realizar una perfecta transducción de la consulta y los índices documentales a una representación interlingüística.

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. La clave de la *web semántica* radica en estructurar el contenido semántico de las páginas *web*, creando así un entorno en el que los agentes de software puedan navegar por Internet con el fin de realizar tareas que satisfagan las necesidades de los usuarios (Eisele y Ziegler-Eisele 2002).

A modo de ilustración, supongamos que queremos ofrecer a los usuarios un buscador online que recupere información de una base de datos que tengamos alojada en nuestro servidor. A diferencia de los buscadores actuales, los cuales se basan en simples procesos de comparación de palabras, nuestra búsqueda tendría una fundamentación semántica. En otras palabras, los documentos serían recuperados según el grado de similitud semántica con respecto a la consulta, la cual podría formularse en una lengua diferente a la de los documentos. En este entorno multilingüe, adoptaríamos un enfoque interlingüístico, donde tanto el texto de la consulta como los documentos de la base de datos consultada se convertirían en estructuras lógicas conceptuales con el fin de comparar su similitud semántica. Esta fase implicaría igualmente la intervención del motor de razonamiento descrito en el aparatado 3, para lo cual las estructuras lógicas

conceptuales se transformarían en esquemas conceptuales en COREL. El motor de búsqueda no operaría directamente sobre los documentos originales sino sobre nuestro propio banco de memoria, el cual estaría formado por archivos XML que representaran cada uno de los documentos de la base de datos. Así, cada archivo XML incluiría información sobre la dirección electrónica del documento original y la estructura lógica conceptual de parte de su contenido. Por ejemplo, para un documento que incluyese el texto (14) podríamos tener el documento (15).

(14) A foot for adjusting the height of a dishwasher or washing machine comprising a threaded shank and a sprocket.

(15) <CLSdoc>

<URL> http://www.freepatentsonline.com/7556227.html</URL>

<CLS>[(do ([+COMPRISE_00 (+FOOT_00-Theme, [$THREADED_SHANK_00-Referent & $SPROCKET_00-Referent])])] CAUSE [BECOME +DIFFERENT_00 ([+BE_01 ([$DISHWASHER_00-Theme ^ $WASHING_MACHINE_00-Theme], +HEIGHT_00-Attribute)])]</CLS>

</CLSdoc>

Una vez obtenidos los índices de similitud basados en la relevancia conceptual entre la estructura lógica conceptual de la consulta[18] y las estructuras lógicas conceptuales en el banco de memoria, el buscador ordenaría decrecientemente las direcciones electrónicas almacenadas el banco de memoria, las cuales permitirían redirigirnos a los documentos originales. De esta forma, demostramos que 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.

[18] La consulta podría tomar la forma de una o varias palabras clave o de una explicación en lenguaje natural—por ejemplo, en relación a (14), podría haber sido "soporte para nivelar la lavadora". En caso de ambigüedad durante el procesamiento de la consulta, el sistema también funcionaría como una agente inteligente que asistiera a los usuarios a refinar sus consultas.

5- Conclusiones

Este capítulo ofrece una descripción selectiva de algunas de las aplicaciones que podemos desarrollar a partir de la noción de ELC, en la cual predomina un destacado basamento ontológico. Como parte fundamental de este giro conceptual, describimos la arquitectura de FunGramKB y sus módulos ontológico y léxico, lo que nos ofrece un marco de representación para entender cómo se construye una ELC. En las dos secciones finales, abordamos el objetivo central de este trabajo: qué impacto tiene la ELC para el desarrollo de aplicaciones en el procesamiento del lenguaje natural. Nos centramos en las dos que hasta la fecha hemos explorado, a saber, un sistema de traducción automática y un sistema de recuperación y extracción de la información en un entorno multilingüe. Sostenemos que el potencial de las ELCs puede abarcar otro tipo de aplicaciones del PLN que, junto con el algoritmo de enlace bidireccional, nos permite validar la adecuación computacional de la GPR.

Agradecimientos

Este trabajo forma parte del proyecto de investigación financiado por el Ministerio de Ciencia y Tecnología, código FFI2008-05035-C02-01.

Referencias

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

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

Bos, J. 2005. Towards wide-coverage semantic interpretation. In *Proceedings of the Sixth International Workshop on Computational Semantics IWCS-6,* Tilburg. 42-53.

Eisele, A. y D. Ziegler-Eisele (2002) “Towards a road map on human language technology: natural language processing”. En *Proceedings of COLING Workshop ‘A Roadmap for Computational Linguistics’*, Taipeh.

Gonzalo Arroyo, J. y M.F. Verdejo Maíllo, 2003, “Recuperación y extracción de información”. *Tecnologías del Lenguaje*. Ed. M.A. Martí Antonín. Universitat Oberta de Catalunya, Barcelona. 157-192.

Google. 2009. Google Translator. http://translate.google.com.

Harman, D., P. Schäuble *et al.*, 1995, “Document retrieval”. *Survey of the State of the*

Art in Human Language Technology. Eds. G. Varile y A. Zampolli. Cambridge University Press, Cambridge. 259-262.

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas y Antonio Zampolli, 2000: "SIMPLE: A general framework for the development of multilingual lexicons", International Journal of Lexicography 13 (4), 249-263.

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

Microsoft. 2009. Microsoft Translator. htttp://www.windowslivetranslator.com/Default.aspx.

Nolan, B., R. Mairal Usón y C. Periñán Pascual. 2009. Natural language applications in an RRG framework. In *Proceedings of the Role and reference Grammar International Conference*. University of California, Berkeley.

Nolan, B. and Y. Salem. 2009. UniArab: an RRG Arabic-to-English machine translation software. In *Proceedings of the Role and reference Grammar International Conference*. University of California, Berkeley.

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

— (2005) "Microconceptual-Knowledge Spreading in FunGramKB", *Proceedings on the Ninth IASTED International Conference on Artificial Intelligence and Soft Computing*, ACTA Press, Anaheim-Calgary-Zurich, pp. 239- 244.

— (2007a) "Cognitive modules of an NLP knowledge base for language understanding", Procesamiento del Lenguaje Natural 39, pp. 197-204.

— (2007b) "Deep semantics in an NLP knowledge base", *12th Conference of the Spanish Association for Artificial Intelligence*. Daniel, Borrajo, Luis Castillo y Juan Manuel Corchado (eds.). Universidad de Salamanca, Salamanca, pp. 279-288.

— (2008) "A cognitive approach to qualities for NLP". Procesamiento del Lenguaje Natural 41, pp. 137-144.

— (2010a) "Ontological commitments in FunGramKB". Procesamiento del Lenguaje

Natural 44, pp. 27-34.

— (2010b) “The Architecture of FunGramKB”, *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Malta, ELRA, pp. 2667-2674.

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

— (2010a) “Enhancing UniArab with FunGramKB”. Procesamiento del Lenguaje Natural 44, pp. 19-26.

— (2010b) “La gramática de COREL: un lenguaje de representación conceptual”. Onomázein 21, pp. 11-45.

— (s.f.) “Constructing the FunGramKB basic conceptual level: the COHERENT methodology”.

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

Salem, Y., A. Hensman and B. Nolan. 2008a. Implementing Arabic-to-English machine translation using the Role and Reference Grammar linguistic model. In *Proceedings of the 8th Annual International Conference on Information Technology and Telecommunication,* Galway, Ireland.

Salem, Y., A. Hensman and B. Nolan. 2008b. Towards Arabic to English machine translation. *ITB Journal*, 17: 20-31.

Salem, Y. and B. Nolan. 2009a. Designing an XML lexicon architecture for Arabic machine translation based on Role and Reference Grammar. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools,* Cairo, Egypt.

Salem, Y. and B. Nolan. 2009b. UniArab: a universal machine translator system for Arabic Based on Role and Reference Grammar. In *Proceedings of the 31st Annual Meeting of the Linguistics Association of Germany*.

Tzoukermann, E., J.L. Klavans *et al.*, 2003, "Information retrieval". *The Oxford Handbook of Computational Linguistics.* Ed. R. Mitkov. Oxford University Press, Oxford. 529-544.

Van Valin, R. 2005. *Exploring the Syntax-Semantic Interface*. Cambridge University Press, Cambridge.

Van Valin, R. and R. LaPolla. 1997. *Syntax: Structure, Meaning, and Function*.

Cambridge University Press, Cambridge.

Vauquois, B. 1968. A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation. *IFIP Congress-68*. 254-260.

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