El módulo léxico de FunGramKB

Fátima Guerra García, Universidad de La Laguna (España)
Elena Sacramento Lechado, Universidad de La Laguna (España)

Índice

Resumen

La base de conocimiento FunGramKB presenta un enfoque conceptualista que permite representar diferentes tipos de conocimiento de forma eficaz. Entre sus niveles principales encontramos el nivel léxico, que almacena información morfosintáctica, pragmática y sobre colocaciones de las unidades léxicas de diferentes lenguas (en su lexicón correspondiente) y que lexicalizan los conceptos existentes en la Ontología. Los diferentes lexicones resultan muy útiles para la extracción de información, traducción o comprensión de textos, pues aportan información completa y detallada sobre cada una de sus unidades léxicas. El propósito de este artículo es describir el nivel léxico de la base de conocimiento, más concretamente el Lexicón, y el proceso de implementación en el editor de FunGramKB Suite.

Palabras clave: FunGramKB, base de conocimiento, lexicón, ontología, Aktionsart, construcciones.

1 Introducción

FunGramKB (Periñán y Arcas, 2007, 2010; Periñán y Mairal, 2009, 2010) es una base de conocimiento léxico-conceptual multipropósito y multilingüe diseñada para su uso en sistemas de PLN, y más específicamente, para aplicaciones que requieran la comprensión del lenguaje natural, ya que FunGramKB integra información sintáctica y semántica muy variada que es de gran ayuda en el desarrollo de dichas aplicaciones. La base de conocimiento comprende tres niveles principales de conocimiento: los niveles léxico y gramatical almacenan el conocimiento lingüístico, mientras que el nivel conceptual almacena el conocimiento no lingüístico compartido por todas las lenguas.

Una de las características principales de FunGramKB es su enfoque conceptualista, lo cual quiere decir que el nivel conceptual, y más concretamente su Ontología, se convierte en el pilar fundamental de toda la arquitectura de la base de conocimiento. Como consecuencia, la creación de dicha base se realiza de manera descendente, i.e. la población de las entradas léxicas requiere la previa modelación

del concepto correspondiente al que están ligadas en la Ontología. De este modo, siempre que los ingenieros del conocimiento hayan creado el concepto +COOK_00 en la Ontología junto con su postulado de significado, marco temático y palabras que lo lexicalizan en diferentes idiomas, los lingüistas o lexicógrafos computacionales podrán rellenar la información concerniente a las unidades léxicas *cocinar,* en el lexicón español, *cook* en el lexicón inglés, *cucinare* en el lexicón italiano, o *cuisiner* en el lexicón francés.

En este contexto, este artículo[1] pretende servir de guía práctica para la formación de ingenieros del conocimiento con el objetivo de impulsar la tarea de poblar los diferentes lexicones de la base de conocimiento. En cuanto a la organización de este trabajo, el capítulo 2 ofrece una pequeña explicación del nivel léxico de FunGramKB seguida de una explicación del formato del editor de dicho nivel y cómo utilizarlo para realizar la búsqueda y subsecuente complementación de las unidades léxicas. El capítulo 3 describe detalladamente los diferentes rasgos de las entradas léxicas de FunGramKB para sustantivos, adjetivos y adverbios, mientras que el capítulo 4 hace lo propio para los verbos. El capítulo 5 explica lo que denominamos información miscelánea, un conjunto que incluye aspectos como el dialecto o el estilo, y que es común a sustantivos, adjetivos, adverbios y verbos. Finalmente, el capítulo 6 presenta algunas conclusiones.

2 El Lexicón de FunGramKB

El Lexicón de FunGramKB almacena información morfosintáctica, pragmática y sobre colocaciones de las unidades léxicas que conforman dicha base de conocimiento. Esta información es compartida con el Morficón, que trata los casos de morfología flexiva que ocurren con las unidades del Lexicón, así como con el resto de niveles en FunGramKB, lo cual garantiza la capacidad informativa y minimiza la redundancia.

En el componente léxico de FunGramKB el conocimiento lingüístico se representa y gestiona de forma eficaz, aportando entradas léxicas que presentan la información de forma más precisa e integrada que los diccionarios tradicionales. El nivel léxico es dependiente de cada lengua, por lo que los lingüistas computacionales deberán desarrollar un Lexicón para cada una de las lenguas incluidas en la base de conocimiento. En FunGramKB, hay actualmente cuatro lexicones (i.e. español, inglés, francés e italiano), aunque en este trabajo nos centraremos en el español y el inglés exclusivamente.

2.1 El editor de FunGramKB

Lo primero que debemos hacer una vez entramos en el editor es introducir la unidad léxica con la que vamos a trabajar en el recuadro de la parte superior izquierda del editor de FunGramKB (véase figura 1). Podemos restringir la búsqueda seleccionando “part of speech” (que hace referencia a la categoría gramatical). Por ejemplo, si sólo queremos trabajar con el sustantivo *jump*, entonces seleccionaremos N (Noun) en “part of speech”; si sólo queremos trabajar con el predicado *jump*, seleccionaremos V (Verb) y si queremos trabajar con todas las clases de palabra, entonces seleccionaremos [no filter].

[1] Remitimos al lector al trabajo realizado por Mairal y Periñán (2009), donde se explican en detalle las bases teóricas del nivel léxico de FunGramKB.

Figura 1: Visión parcial del editor de FunGramKB para los lexicones español e inglés.

A continuación, el ordenador nos mostrará los resultados de la búsqueda en un recuadro que aparece inmediatamente debajo. Si la unidad léxica en cuestión tiene un único significado almacenado en la base de conocimiento obtendremos un solo resultado. Así, para la búsqueda de la unidad léxica *jump,* el ordenador sólo nos muestra la siguiente entrada: jump [+JUMP_00] (entre corchetes se encuentra el concepto de la Ontología al que está ligada la unidad léxica en cuestión).

Por otro lado, si la palabra es polisémica estará lexicalizando a más de un concepto en la Ontología y el ordenador nos mostrará tantos resultados como sentidos tenga la palabra. Por ejemplo, en español *pegar* tiene dos sentidos asociados ontológicamente a dos conceptos distintos: por un lado *pegar* significa "maltratar con golpes" y por otro "adherir una cosa a otra". De la misma manera, si introducimos en el lexicón inglés la unidad léxica *cook*, seleccionando la opción [no filter], veremos todos los sentidos de esa unidad léxica que han sido almacenados hasta la fecha en la base de conocimiento, que son:

Cook [+COOK_00] = *cocinar*
Cook [$COOK_00] = *cocinero*

Lo mismo ocurre con *run*:

run [+RUN_00] = *correr*
run [+OPERATE_00] = *poner en marcha, operar, encender*
run [+MOVE_00] = *fluir, correr*

Esto quiere decir que la unidad léxica tiene tres sentidos diferentes y como consecuencia, ha sido anotada en tres conceptos distintos por los ingenieros del conocimiento trabajando en la Ontología.

Al seleccionar uno de los conceptos se abrirá la "ficha" sobre la unidad léxica. En la parte superior derecha veremos la información conceptual del concepto al que está ligada. Por ejemplo, para el primer caso (run [+RUN_00]) la información conceptual almacenada en la ontología es la siguiente:

LEXICAL UNIT:	run
CONCEPT:	+RUN_00
THEMATIC FRAME:	(x1)Agent (x2: +HUMAN_00 ^ +ANIMAL_00)Theme (x3: +GROUND_00)Location (x4)Origin (x5)Goal
MEANING POSTULATE:	+(e1: +WALK_00 (x1)Agent (x2)Theme (x3)Location (x4)Origin (x5)Goal (f1: +FAST_00)Speed)
DESCRIPTION:	move fast by using one´s feet, with one foot off the ground at any given time

Toda esta información, que incluye el marco temático, el postulado de significado y una descripción de diccionario, ayuda a comprender mejor el sentido del que nos estamos ocupando, por lo que debemos tenerla siempre presente cuando comencemos a anotar información sobre la unidad léxica.

Si seleccionamos el segundo caso (run [+OPERATE_00]) veremos que la información conceptual que se nos muestra es otra:

LEXICAL UNIT:	run
CONCEPT:	+OPERATE_00
THEMATIC FRAME:	(x1: +HUMAN_00)Theme (x2: +ARTEFACT_00)Referent
MEANING POSTULATE:	+(e1: +USE_00 (x1)Theme (x2)Referent (f1)Instrument (f2: (e2: +DO_00 (x2)Theme (x3)Referent))Purpose)
DESCRIPTION:	to use and control a machine or equipment

En este caso estaríamos trabajando con el predicado verbal *run* con el sentido “to use and control a machine”, es decir, “poner en marcha, encender”; por lo tanto, todos los datos que introduzcamos a continuación deben referirse a este sentido de la palabra obviando los otros dos.

Debajo de la información conceptual encontramos la ficha de la unidad léxica en la que guardamos información específica sobre la misma y que pasaremos a explicar en los siguientes apartados.

3 Información para sustantivos, adjetivos y adverbios

En FunGramKB, todos los sustantivos, adjetivos y adverbios *(entities* y *qualities)* serán almacenados en su Lexicón correspondiente (i.e. español, inglés, francés o italiano). La información que debe ser completada y almacenada en este módulo se resume a continuación.

3.1 Información morfosintáctica del sustantivo

La información morfosintáctica sobre el sustantivo incluye diversos aspectos como variantes gráficas, abreviaturas, principales constituyentes, número o contabilidad. Debemos anotar esta información en las casillas correspondientes.

Las **variantes gráficas** son palabras que presentan una doble grafía, es decir, palabras que se escriben de forma diferente aunque hacen referencia al mismo término. Este es el caso de *behavior-behaviour* en inglés o *septiembre-setiembre* en español. Generalmente, estas variantes se emplean de forma diferente en las

variedades dialectales de una lengua o su uso varía según el contexto en el que se utilicen.

Las **abreviaturas** incluyen ejemplos como *television-TV*, *advertisement-ad/advert,* en inglés, o *fotografía-foto* en español. En la mayoría de los casos, las abreviaturas se usan de forma frecuente en el uso diario de la lengua. Además, en la casilla destinada para este uso, anotaremos también acrónimos y siglas, como *radar* (Radio Detection and Ranging) o *CD* (Compact Disc).

En el caso de encontrarnos ante un **modismo o giro idiomático**, anotaremos en la casilla "Phrase Constituents: Head" cuál de sus componentes funciona como núcleo o constituyente principal. Por ejemplo, la expresión idiomática inglesa *"as mad as a March hare"* estaría ligada al concepto $CRAZY_00 en la Ontología, y su núcleo o componente principal sería el adjetivo *mad,* por lo que deberemos anotarlo. De igual forma, para su equivalente en español *("loco como una cabra"),* anotaríamos el adjetivo *loco* como constituyente principal.

En cuanto al **número**, podremos indicar si el sustantivo es o no dual. La dualidad hace referencia a los casos en los que la oposición singular-plural implica la adición de marcadores morfológicos, como la terminación *–s* en *cereza-cerezas* o *–es* en *potato-potatoes*. En este caso, marcaremos "Dual: Regular". Sin embargo, si la formación de plural implica un cambio vocálico en la raíz o la adición de terminaciones diferentes (como *–en,* por ejemplo), entonces marcaremos "Dual: Irregular". Se incluyen en este grupo ejemplos como *woman-women* o *tooth-teeth.*

Por el contrario, las unidades léxicas no-duales expresan su número a través de marcadores sintácticos, pues se requiere la adición de artículos o numerales para saber si el sustantivo es singular o plural (p.ej. *deer, series* o *sheep* en inglés; *viernes, crisis* o *tesis* en español). En este caso, indicaremos "Common", pues se emplea una unidad léxica común para singular y plural. Además, podemos encontrarnos con casos de sustantivos en *singularia tantum*, es decir, aquellos que son usados únicamente en singular, como *dust* en inglés o *salud* en español, o con casos de *pluralia tantum,* que hacen referencia a aquellos sustantivos que presentan una morfología plural, y por lo tanto, rigen plural aunque se refieran a un solo objeto (p.ej. *trousers, tijeras* o *víveres).*

Indicaremos también la **contabilidad** marcando la opción correspondiente, esto es, contable *(countable),* incontable *(uncountable/mass nouns)* o dual. Los sustantivos contables señalan entidades que se pueden contar, por lo que podrán combinarse con numerales o cuantificadores (p.ej. *butterfly, tourist, libro* o *pez).* Los sustantivos incontables señalan entidades que no se pueden contar, pero que pueden ser medidas o cuantificadas (p.ej. *furniture, money, aire* o *harina).* Finalmente, los duales hacen referencia a aquellos sustantivos que pueden comportarse como contables o incontables, por lo que su cuantificación puede expresar tanto cantidad como número (p.ej. *mucho pollo-muchos pollos).*

En el lexicón español, además, encontraremos una casilla adicional para seleccionar el **género** del sustantivo con el que estemos trabajando. De esta forma, seleccionaremos "Masculine" para los sustantivos masculinos *(mapa, libro),* "Feminine" para los femeninos *(plata, revista),* y "Ambiguous" para aquellos que admiten ambos géneros *(el mar/la mar).* Además, marcaremos "Dual: Regular" si la unidad léxica en cuestión hace referencia a una entidad sexuada, y por tanto, puede ser tanto masculina como femenina. Este es el caso de aquellos sustantivos cuya forma femenina se marca a través de la adición de un morfema de género (i.e. *gato/gata, niño/niña).* Por otra parte, si se utilizan palabras diferentes para diferenciar el sexo, seleccionaremos "Dual: Irregular" (i.e. *caballo/yegua).* Finalmente, seleccionaremos "Common" cuando se cambia el género gramatical del determinante pero se mantiene la misma unidad léxica *(el joven/la joven, el líder/la líder).*

3.2 Información morfosintáctica del adjetivo y del adverbio

La información morfosintáctica sobre el adjetivo y el adverbio también incluye aspectos como variantes gráficas, abreviaturas y principales constituyentes. Además de esta información, debemos indicar la categoría, el grado y la posición adjetival.

Para la **categoría**, seleccionaremos "a" si la unidad léxica puede funcionar como adjetivo o adverbio (p.ej. *fast),* "adj" si se trata de un adjetivo únicamente (p.ej. *sweet* o *delicado),* y "adv" si se trata de un adverbio exclusivamente (p.ej. *lately, lovely, ahora* o *lejos).*

El **grado** hace referencia a la forma de expresar la intensidad relativa de los adjetivos. De esta forma, seleccionaremos "Regular-Inflectional" si las formas comparativa y superlativa se construyen añadiendo sufijos *(tall – taller – the tallest);* "Regular-Periphrastic" si el comparativo y superlativo se forman utilizando palabras como *more* y *most* (en inglés) o *más* (en español) en lugar de sufijos *(expensive - more expensive - the most expensive; grande – más grande – el más grande);* "Irregular" si la gradación implica un cambio morfológico *(far – further – the furthest, bueno – mejor – óptimo),* o "None" si la unidad léxica no admite gradación (p.ej. *eternal, principal, absoluto).* Es interesante señalar que las formas irregulares se almacenarán en el Morficón correspondiente.

Además, añadiremos la información relativa a la **posición** más común de la unidad léxica dentro del sintagma. Los adjetivos atributivos directamente califican al sustantivo otorgándole un atributo (p.ej. *auto rojo, película terrorífica),* mientras que los predicativos son aquellos que van ligados al nombre mediante un verbo copulativo (p.ej. *la película resulta interesante).* En el caso de que las dos opciones sean posibles, marcaremos "Dual". En español, los adjetivos en posición atributiva recibirán mayor especificación dependiendo de su función: pre-modificadores (p.ej. *buen hombre)*, post-modificadores (p.ej. *lápiz rojo)* o (pre-/post-)modificadores (p.ej. *nieve blanca / blanca nieve),* por lo que el lexicón español presenta una casilla adicional para almacenar esta información.

4 Información para verbos

A continuación, explicaremos paso por paso la información que debe ser almacenada sobre predicados verbales.

4.1 Información morfosintáctica del verbo

En primer lugar, anotamos información morfosintáctica sobre el verbo que incluye aspectos como variantes gráficas, abreviaturas, principales constituyentes, tipo de conjugación verbal y pronominalización.

Las **variantes gráficas**[2] incluyen ejemplos como *fulfill-fulfil* o *realize-realise.*

Las **abreviaturas** son mucho más comunes para sustantivos que para verbos, no obstante la entrada de los verbos contempla la posibilidad de incluir esta información.

En el caso de que la unidad léxica con la que estemos trabajando sea un *phrasal verb* debemos anotar cuál es su **núcleo** y cuál su **partícula**. Por ejemplo, para *show up, run away* o *look up* marcaremos *show, run* y *look* como los núcleos o constituyentes principales (*heads)* de la expresión, mientras que *up* y *away* son las partículas adverbiales o preposicionales. Si estamos tratando con un ejemplo de *phrasal verb* intransitivo, aquellos que no pueden ir seguidos de un objeto, dejaremos la casilla "*detachable"* sin marcar. Sin embargo, para los *phrasal verbs* transitivos que

[2] Los verbos con variantes gráficas son menos frecuentes que los sustantivos, especialmente en la lengua española.

sí pueden ir seguidos de objetos, pudiendo estos aparecer tanto después como antes de la partícula, marcaremos la casilla "*detachable"*; por ejemplo, seleccionaremos "*detachable"* para la partícula *up* de *look up* (*I looked the number up in the phone book = I looked up the number in the phone book*). Por otro lado, en el lexicón español sólo existe la casilla correspondiente al núcleo de la expresión (*head)* para casos como *meter la pata,* donde sólo debemos anotar que el núcleo de la expresión es *meter.*

Dentro de las características del **paradigma verbal** anotaremos el tipo de conjugación del verbo. Son verbos regulares aquellos que mantienen inalterado su lexema o raíz a lo largo de su conjugación, como por ejemplo *comer, vivir, arrive* o *dance.* Los verbos irregulares presentan variaciones o modificaciones en el paradigma a la hora de conjugar el tiempo y/o modo verbal, por ejemplo los verbos *tener, hacer* (en español)*, eat* o *come* (en inglés). Por último, también existen casos en inglés de verbos con los dos usos, regular e irregular, por ejemplo *learn-learnt/learned-learnt/learned,* para los que seleccionaremos la opción *(Ir)regular.* Otro aspecto que podemos indicar es si el paradigma verbal tiene restricciones, lo que en español se denomina "verbos defectivos" y que se caracterizan por tener su paradigma verbal incompleto al no existir la conjugación de algunos de sus tiempos o personas. Algunos ejemplos del español son el verbo *abolir,* que no tiene todos los tiempos verbales, u *ocurrir, suceder, llover, nevar, etc.,* que no permiten todo tipo de sujetos y por lo tanto, la persona del paradigma verbal se ve limitada.

Por último, indicaremos los valores de **pronominalización**[3] del verbo, que contemplan aquellos fenómenos que suponen variaciones clíticas de un lema, p.ej. reflexividad y reciprocidad.

Por un lado, los valores de **'reflexividad'** son:

(i) *Never* (nunca reflexivo): el verbo no admite el pronombre "se", p.ej. *parir.*
(ii) *Always* (siempre reflexivo): el pronombre reflexivo "se" es obligatorio, p.ej. *jactarse.*
(iii) *Optional:* el verbo puede marcarse como reflexivo con el pronombre clítico "se", pero la presencia del "se" es opcional y no implica un cambio en la denotación del verbo, p.ej. *ir(se)*.
(iv) *Grammatical:* variante contextual del "se" tradicionalmente concebida como un mecanismo gramatical. El pronombre reflexivo "se" aparece como un uso variante del verbo que afecta su transitividad canónica (i.e. el número de argumentos) según alguno de los siguientes criterios:
  (a) Los verbos transitivos se marcan reflexivamente para indicar la identidad de dos variables argumentales en la plantilla léxica, p.ej. *Se miró en el espejo.*
  (b) Se introduce "se" para dejar sin expresión explícita en la sintaxis al argumento Tema. Esto ocurre en:
  - oraciones pasivas: *Muchas pirámides se construyeron en el México antiguo,*
  - oraciones decausativas: *El vaso se rompió,*
  - 'reflexivos indeterminados' referencialmente (Robertson y Turley, 2003): *Por aquí se come mucho helado.*

Por otro lado, los valores de **'reciprocidad'** son:

(i) *Never:* no se puede usar un pronombre recíproco, p.ej. *beber.*
(ii) *Grammatical:* se usa el recíproco para indicar que las entidades implicadas (normalmente el sujeto es plural) ejecutan el evento sobre cada una. La consecuencia es que se reduce el número de argumentos del verbo, p.ej. *casarse.*

[3] Para más información sobre el tratamiento de la pronominalización en los lexicones de FunGramKB véase Guerra García y Sacramento Lechado (2011).

Los lexicones de FunGramKB tienen una arquitectura con una orientación sensible a los sentidos de las palabras. Esto condiciona el tratamiento de la pronominalización, especialmente en aquellos casos en los que la presencia del clítico altera el significado del verbo. Por ejemplo, las unidades léxicas *acordar* y *acordarse* están ligadas a conceptos diferentes en la Ontología (i.e. *acordar* está ligada al concepto +AGREE_00 y *acordarse* está ligada a +REMEMBER_00), por lo que dos entradas léxicas diferentes son creadas. Este hecho determina que *acordar* nunca es reflexivo, mientras que *acordarse* siempre es reflexivo.

Una vez hayamos tomado una decisión sobre la reflexividad y la reciprocidad de la unidad léxica con la que estemos trabajando, es aconsejable comprobar la veracidad de nuestras conclusiones en un corpus, sobre todo con los casos más dudosos. Algunos corpora recomendados en la página web de FunGramKB son el BNC (British National Corpus)[4] o COCA (Corpus of Contemporary American English)[5] para el inglés y CREA (Corpus de Referencia del Español Actual)[6] o Corpus del Español de Mark Davies[7] para el español.

4.2 Gramática nuclear

En esta sección introducimos información relacionada con la gramática nuclear del verbo, la cual está inspirada en el modelo de la GPR (Gramática del Papel y la Referencia). Así, información como el *Aktionsart* del verbo, las variables, papeles temáticos y construcciones, entre otras, son anotadas en este apartado.

Siguiendo la clasificación de las clases de verbos de la GPR[8] y haciendo uso de los siete tests lingüísticos[9], indicaremos el ***Aktionsart*** del verbo. Pueden apuntarse tres *Aktionsarten* como máximo: por ejemplo, para *run* se han seleccionado actividad y realización activa, y para *push* actividad causada y realización activa causada.

A continuación, seguiremos con información relacionada con la plantilla léxica. Debemos apuntar el número máximo de **variables** del predicado, como por ejemplo tres variables (x, y, z) para el verbo *enviar/send.* En los casos en los que seleccionemos más de un *Aktionsart*, anotaremos el número de variables del mayor. Por ejemplo, el verbo *soplar/blow* es una actividad, como en *He was blowing onto her coffee while reading the newspaper,* y también una realización activa, como en *He blew the candles in a second,* así que el número máximo de variables sería dos (x, y), los correspondientes a la realización activa y no la única variable (x) de su uso como actividad.

El siguiente valor a rellenar es el de los **macrorroles o macropapeles.** Los macrorroles son dos, actor y padecedor, y son en términos generales el sujeto y objeto lógico en términos de lógica formal, lo cual sirve para predecir el comportamiento sintáctico de los argumentos y, por lo tanto, la transitividad del predicado. Siguiendo de nuevo la teoría de la GPR[10], un predicado puede tener:

- 0 macrorroles [MR0] o atransitivo, por ejemplo el verbo *snow* o *rain.*
- 1 macrorrol [MR1] o intransitivo, por ejemplo el verbo *die.*
- 2 macrorroles [MR2] o transitivo, por ejemplo *eat/comer* (realización activa: *"Se comió un trozo de pizza"*), *kill* (realización causativa) o *send* (logro causativo).

[4] http://www.natcorp.ox.ac.uk/
[5] http://corpus.byu.edu/coca/
[6] http://corpus.rae.es/cordenet.html
[7] http://www.corpusdelespanol.org/
[8] Van Valin y LaPolla (1997) y Van Valin (2005).
[9] Van Valin (2005).
[10] Para más información sobre la transitividad en RRG, tanto transitividad-S(intáctica) como transitividad-M(acrorole), véase Van Valin y LaPolla (1997:150).

Debemos anotar siempre cuántos macrorroles toman los predicados verbales con los que trabajemos (i.e. MR0, MR1 o MR2). Además, debemos marcar cuál de los argumentos del verbo actúa como padecedor[11]; para ello se siguen los siguientes criterios:

- Si el verbo es MR0 no marcamos [U], ya que no se ha tomado macrorrol alguno. *P.ej. nevar, llover* [U = no value selected]
- Si el verbo es MR1:
    - Si el verbo es una actividad o un semelfactivo (es decir, dinamismos) el verbo sólo toma un macrorrol (MR1) y este es siempre por defecto actor (x = actor). Por lo tanto, no procede marcar valor alguno al [U]. P.ej. *bailar/dance* → [MR1][U = no value selected] o *toser/cough* → [MR1][U = no value selected]
    - Si el verbo no es una actividad (es decir, no es un dinamismo) el macrorrol será siempre padecedor (x = padecedor). Esto ocurre con los estados, logros y realizaciones. P.ej. *derretir(se)* → [MR1][U = x]
- Si el verbo es MR2 debemos especificar qué argumento del verbo actúa como padecedor:
    - Si el verbo es causativo marcaremos [U = y], ya que el argumento x del verbo suele actuar como actor mientras que el argumento y padece la acción. P.ej. *empujar/push* → [MR2 U=y] o *abrir/open* → [MR2 U=y]
    - Los verbos de cambio de posesión *take, get, buy, steal, rob* o *donate,* no participan en la alternancia de dativo, por lo que la asignación del padecedor al tercer argumento tiene que especificarse en su entrada léxica como [U=z]. P.ej. *donate* → [MR2][U = z]

Los rasgos de gramática nuclear son los que más información aportan a la hora de hacer un análisis de lingüística comparada entre diferentes lenguas. Por ejemplo, mientras en la Ontología el concepto básico +LIKE_00 almacena información conceptual en su marco temático y postulado de significado que es común para todas las lenguas, es en el lexicón donde descubriremos las particularidades e idiosincrasias de los diferentes verbos que lexicalizan ese concepto en diferentes lenguas. Así, en el lexicón inglés comprobaremos que el verbo *like* es MR2 U=y mientras que, en el lexicón español, *gustar* es MR1 U=y, siendo ese el motivo por el que en inglés se dice *Ana likes John* mientras que en español se dice *A Ana le gusta Juan* y no **Ana gusta Juan*[12].

A continuación almacenaremos la información relativa a la ***proyección de los papeles temáticos***, la cual es crucial para ver cómo se produce el enlace entre la información conceptual del marco temático del concepto en el nivel cognitivo y la información más específica de los predicados en el lexicón. Esto se consigue asignándole a las variables del predicado en el lexicón el correspondiente papel temático de FunGramKB. Los papeles temáticos, cuyo inventario es muy limitado (i.e. *agent, theme, referent, location, origin, goal* y *attribute*), son participantes prototípicos que intervienen en la dimensión cognitiva a la que está vinculado el predicado con el que estemos trabajando. Utilizando los mismos papeles temáticos en los conceptos en la Ontología y en las unidades léxicas en el nivel léxico garantizamos el intercambio de información entre estos dos niveles de la base de conocimiento.

[11] El término padecedor es la traducción del original *undergoer* (Van Valin y LaPolla (1997), Van Valin (2005)), siendo esta la razón por la que el valor de padecedor en la entrada léxica se marca con una U.

[12] La construcción *A Ana le gusta Juan* muestra que el verbo *gustar* tiene un solo macropapel, i.e. el padecedor *Juan,* ya que el otro argumento lleva preposición, i.e. *a Ana,* y los argumentos con macropapel nunca se realizan con preposición.

Por ejemplo, en el caso del verbo *run* hemos seleccionado como *Aktionsarten* actividad y realización activa. Para el *Aktionsart* de actividad tendríamos sólo una variable (x) pero para su uso como realización activa habría dos variables (x, y) y, por tanto, como debemos marcar el número máximo de variables, apuntamos dos (x, y). El número de macrorroles que toma el predicado sería MR1 (i.e. un verbo intransitivo) y no asignamos padecedor (i.e. U = 0) porque en los verbos de movimiento que se convierten en realización activa, la "y" corresponde al GOAL/DESTINO, que es una excepción en la asignación de macrorroles. Por último, especificamos que la variable "x" corresponde al papel temático Tema (i.e. entidad que cambia su lugar o posición) en FunGramKB y que en el uso como realización activa la "y" sería el Destino (i.e. lugar hacia el que se mueve la entidad).

En el caso de *cocinar,* se trata de una actividad y realización activa con un número máximo de dos variables (x, y) con dos macrorroles MR2, lo cual quiere decir que es un verbo transitivo en el que "x" es el Tema (i.e. entidad que crea otro entidad) e "y" es Referente (i.e. entidad creada).

En la siguiente casilla del editor anotaremos también si alguna de las variables suele aparecer con alguna **preposición** prototípica, por ejemplo *run (to/towards), ram (into), rotate/revolve (around)*, o si presenta alguna **colocación** muy frecuente. Las colocaciones son combinaciones de dos o más palabras que a menudo aparecen juntas, por ejemplo *ahorrar (y = dinero, tiempo), pursue (y = policies, goals, interests, objectives).* De nuevo, la búsqueda en corpora suele ser de gran ayuda para encontrar las preposiciones y colocaciones más frecuentes.

Como último aspecto de la gramática nuclear señalaremos las **construcciones**[13] en las que la unidad léxica puede participar. Para esta tarea contamos con la inestimable ayuda de la obra de Levin (1993). Por ejemplo, el predicado *walk* parece participar en las siguientes construcciones: *resultative construction (Cathy walked herself to exhaustion), induced action alternation (Claire walked the dog down the street/The dog walked down the street)* y *preposition drop alternation (She walked across the town/She walked the town).*

5 Información miscelánea

La información miscelánea incluye aspectos como dialecto, estilo, dominio, ejemplos concretos y traducciones. Esta información es común para todas las clases de palabras, i.e. tanto las entradas léxicas de sustantivos como las de adjetivos, adverbios o verbos demandan rellenar esta información. A continuación, se describe en detalle cada uno de estos rasgos.

Para el **dialecto**[14], en el caso del inglés, seleccionaremos "standard" si la unidad léxica se usa de forma general en cualquier variedad dialectal (p.ej. *elephant, run, go, paint),* o British/American, dependiendo de la variedad dialectal en la que se emplee dicha unidad. Por ejemplo, el sustantivo *pavement* se emplea en inglés británico, mientras que *sidewalk* es su equivalente en inglés americano. Lo mismo ocurre con unidades como *biscuit/cookie* o *boot/trunk.* Algunos ejemplos de predicados verbales pertenecientes a un dialecto específico podrían ser *broil* o *charbroil*, que se usan específicamente en inglés americano, mientras que *grill* se emplea en inglés británico. El lexicón español no tiene opción de dialecto hasta la fecha, por lo que sólo se puede seleccionar el valor estándar (*standard)*.

En cuanto al **estilo**, seleccionaremos la opción adecuada de entre las posibilidades que se ofrecen: *common, formal, informal, literary, slang* (común, formal,

[13] Véase Guerra García y Sacramento Lechado (2011) para un análisis más detallado de las construcciones en FunGramKB.

[14] Las variaciones dialectales, al igual que las variantes gráficas, suelen producirse más frecuentemente con sustantivos.

informal, literario o jerga). De esta forma, en palabras como *elucidation* o *grievous* se seleccionarán las opciones "formal" y "literario" respectivamente, mientras que en el caso de *cool* ("fashionable, interesting") o *rechoncho* marcaremos la opción "informal". En cuanto a los verbos, tanto el *Longman Dictionary of Contemporary English* como el *Cambridge Advanced Learner's Dictionary* definen el verbo *weep* en inglés como un sinónimo formal o literario de *cry* y lo mismo ocurre con *entomb,* que es definido como sinónimo formal de *bury*. Algunos ejemplos en español serían *atestiguar* ("declarar, afirmar") o *sobreseer* ("cesar, desistir") para los cuales seleccionaríamos la opción "formal", o *jorobar* ("molestar, fastidiar") que tendría un uso informal. En muchas ocasiones, los diccionarios nos indican el uso especializado o concreto de algunos sustantivos o verbos mediante comentarios tales como *formal*, *specialized* o *technical, slang,* delante de la definición.

El **dominio** hace referencia al área de conocimiento o materia con la que la unidad léxica está relacionada, p.ej. arqueología, arte, literatura, comercio, deporte, gastronomía, derecho o economía. Así, para *fricassee* o *guiso*, por ejemplo, marcaremos "alimentation" y para *courtroom* o *sentencia* seleccionaremos "law". Algunos verbos pertenecientes a dominios específicos pueden ser, por ejemplo, *copulate* o *copular,* para los que marcaríamos el dominio "sexuality" y *chisel* o *cincelar,* para los que podríamos marcar el dominio "artisanship". Si consideramos que la unidad léxica puede pertenecer a varios dominios podemos marcar hasta tres de ellos. Si, por otro lado, las unidades léxicas con las que estamos trabajando se usan en un contexto general y forman parte del vocabulario básico y cotidiano de la lengua, marcaremos "factotum".

A continuación, se anotarán dos **ejemplos** extraídos de un corpus en los que el significado de la unidad léxica quede representado de forma clara y precisa. Generalmente, haremos uso del BNC (British National Corpus) y del COCA (Corpus of Contemporary American English) para el inglés, y del CREA (Corpus de Referencia del Español Actual) o el Corpus del Español (Mark Davies) para el español.

Debemos elegir ejemplos adecuados en extensión y, en el caso de los verbos, que representen alguna de las construcciones típicas en las que participe o que muestren su *Aktionsart* de manera clara y sencilla. Por ejemplo, para los verbos *enviar* (en el lexicón español) y *send* (en el lexicón inglés) anotaremos algún ejemplo que muestre las tres variables máximas que pueden tomar:

- *They can quite easily send you a postcard* (BNC/BNL/W_religion) o
- *[L]os miembros del Tribunal dispondrán de un mes para enviar a la Comisión de Doctorado un informe individual* (CREA/Misc/1999).

De la misma manera, para el verbo *chase* es recomendable buscar un ejemplo que refleje el carácter de actividad del verbo y las dos variables prototípicas que suele tomar (x, y) como por ejemplo:

- *A couple of kids were chasing a stray dog*. (COCA/1999/FIC/Analog).

También es recomendable que si hemos anotado anteriormente alguna preposición prototípica para alguna de las variables, alguno de los ejemplos que elijamos sea una muestra de ello. Por ejemplo, para la unidad léxica *speak* con tres variables (x, y, z) se han marcado como preposiciones frecuentes y = *about,* z = *with, to*; por lo tanto, sería conveniente seleccionar ejemplos como:

- *Mikhail Gorbachev, has spoken publicly about the importance of introducing democracy into the area of production* (BNC/EVP/W_ac_polit_law_edu), o
- *He is resolved to speak with him though he kicks him outdoors* (BNC/CFF/W_biography).

Lo mismo ocurriría con las colocaciones. Por ejemplo, si hemos anotado que el predicado verbal *ahorrar* suele aparecer con la colocación *dinero,* intentaremos en la medida de lo posible buscar un ejemplo que contenga esa colocación.

Finalmente, elegiremos la **traducción** por defecto de la unidad léxica con la que estemos trabajando en las casillas correspondientes para cada lengua (p.ej. inglés, español, francés o italiano). Por ejemplo, seleccionaremos las unidades léxicas *sentimiento* en español y *feeling* en inglés como traducciones por defecto para el concepto básico +FEELING_00 en la Ontología, y para el verbo *saltar* en español seleccionaremos *jump* como traducción por defecto para el inglés, *sauter* para el francés y *saltare* para el italiano. Si el Editor no nos facilita una lista de posibles sinónimos entre los que elegir la traducción por defecto para alguna de las lenguas de la base de conocimiento, significa que esa unidad léxica aún no ha sido anotada en la Ontología lexicalizando algún concepto. Hay que recordar que la base de conocimiento está conceptualmente orientada y que se construye siguiente un método descendente.

6 Conclusiones

En la Ontología del nivel conceptual de FunGramKB podemos observar cómo se organiza el conocimiento semántico en una jerarquía taxonómica a través de conceptos. Mientras el conocimiento semántico almacenado para cada concepto (i.e. marco temático y postulado de significado) es común a todas las lenguas comprendidas en la base de conocimiento, es decir, es compartido por todas las unidades léxicas que lexicalizan el concepto en inglés, español, italiano o alemán entre otros, es en el nivel léxico donde las idiosincrasias de cada una de las lenguas son analizadas y almacenadas.

La representación completa y detallada de la información almacenada en el nivel léxico aporta riqueza lingüística a las unidades léxicas correspondientes a los conceptos de la Ontología, por lo que el conocimiento multilingüe se gestiona de forma eficaz, afianzando la base de conocimiento.

Agradecimientos

Este trabajo forma parte del proyecto de investigación financiado por el Ministerio de Ciencia e Innovación CONSTRUCCIÓN DE UNA BASE DE DATOS LÉXICA Y CONSTRUCCIONAL INGLÉS-ESPAÑOL EN EL NIVEL DE GRAMÁTICA NUCLEAR, código FFI2008-05035-C02-02, y la realización del mismo se enmarca en la investigación subvencionada por el Ministerio de Ciencia e Innovación (Beca de Formación de Personal Investigador BES-2009-017546, convocatoria 2009) y por la entidad bancaria CajaCanarias (Beca CajaCanarias de Investigación para Posgraduados, convocatoria 2010).

Referencias

Guerra García, Fátima y Sacramento Lechado, Elena (2011): “Tutorial de FunGramKB Suite para los Lexicones”. Informe Técnico. Disponible online en: <http://www.fungramkb.com/workbench.aspx>

Levin, Beth. (1993): *English Verb Classes and Alternations.* Chicago and London: The University of Chicago Press.

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

Mairal Usón, Ricardo y Ruiz de Mendoza Ibáñez, Francisco J. (2008): "An Overview of the Lexical Constructional Model: levels of description and subsumption processes". En: 26th *International Conference of the Spanish Society for Applied Linguistics* (AESLA). University of Almería, Spain, 3-5 April.

Periñán Pascual, Carlos y Arcas Túnez, Francisco (2006): "Reusing computer-oriented lexica as foreign-language electronic dictionaries". *Anglogermánica Online* 4, 69-93.

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

Periñán Pascual, Carlos y Arcas Túnez, Francisco (2010): "The Architecture of FunGramKB". En: *7th International Conference on Language Resources and Evaluation*, Malta, pp. 17-23.

Periñán-Pascual, Carlos y Mairal Usón, Ricardo. (2010): "La gramática de COREL: un lenguaje de representación conceptual". *Onomázein* 21 (2010/1), pp. 11-45.

Robertson, John S. y Turley, Jeffrey S. (2003): "A Peircean analysis of the American-Spanish clitic pronoun system". *Semiotica* 145 (1), 21-70.

Ruiz de Mendoza Ibáñez, Francisco J. y Mairal Usón, Ricardo (2007): "Levels of semantic representation: where lexicon and grammar meet". *Interlingüística* 17, pp. 26-47.

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

Van Valin, Robert D. (2005): *The Syntax-Semantics-Pragmatics Interface: An Introduction to Role and Reference Grammar.* Cambridge: Cambridge University Press.

Apéndice 1. Valores en las entradas léxicas de los verbos en FunGramKB.

	VERBOS	
	Inglés	Ejemplo: *run*
1. Información básica[15]		
1.1. Palabra	UL (Unidad Léxica)	Run
1.2. Índice	(Serves to arrange the various sense of a UL)	02
1.3. Lengua	English	English
2. Morfosintaxis		
2.1. Variantes gráficas	Anotar	
2.2. Abreviaturas	Anotar	
2.3. Constituyentes		
2.3.1. Núcleo	Anotar	
2.3.2. Partícula	Anotar	
2.4.Categoría	Verb	Verb
2.5. Paradigma verbal	Regular / Irregular / (Ir)regular	Irregular
2.5.1. Restricciones en el paradigma	Sí / No	No
2.6.Pronominalización:		
2.6.1. Reflexividad	Never / optional / grammatical	Never
2.6.2. Reciprocidad	Never / optional / grammatical	Never
3. Gramática Nuclear LCM		
3.1. *Aktionsart*	Activity/State/Achievement/Accomplishment/ Semelfactive/Active Accomp./Caus. activity/Caus. State/Caus. Achvm./Caus. Accomp./Caus. Active Accompl./Caus. semelfactive	Actividad / Realización activa
3.2. Plantilla léxica		
3.2.1. Variables	x, y, z	x, y
3.2.2. Rasgos idiosincrásicos:		
MR	MR0 / MR1 / MR2	MR2
Padecedor	U = no value selected / x / y / z	U = no value selected
3.2.3. Proyección de los papeles temáticos[16]	x = no value selected / Agent / Theme / Referent / Goal / etc. y = no value selected / Agent / Theme / Referent / Goal / etc. z = no value selected / Agent / Theme / Referent / Goal / etc.	x = Tema y = Goal
3.2.4. Preposiciones	x = seleccionar del desplegable/ y = seleccionar del desplegable / z = seleccionar del desplegable	y = to, towards
3.2.5. Colocaciones	x = anotar / y = anotar / z = anotar	
3.3. Construcciones	Seleccionar de los desplegables	Induced Action Alternation
4. Miscelánea		
4.1. Dialecto	American / British / Standard	Standard
4.2. Estilo	Common / formal / informal / literary / slang	Common
4.3. Dominio	Arts / biology / medicine / transport /agriculture / law / religion /sexuality / linguistics / alimentation / fashion / FACTOTUM / etc. (seleccionar del desplegable)	Factotum
4.4. Ejemplos	Anotar	*When the boat was under way she looked back and there were the children running in the dunes […]* [BNC/FPF/W_fict_prose] *Frightened by the stranger he ran to the kitchen […]* [BNC/AMB/W_fict_prose]
4.5. Traducción	Seleccionar	Spanish = *correr* French = *courir* Italian = *correre* German = *laufen*

[15] Esta información es facilitada automáticamente.

[16] Los papeles temáticos a seleccionar que aparecen en el desplegable de la proyección de los papeles temáticos dependerán de la información conceptual que el razonador recupere sobre el concepto al que está ligada la unidad léxica.