Construcción de una base de datos de colocaciones léxicas

Margarita Alonso Ramos (Universidade da Coruña, lxalonso@udc.es)
Begoña Sanromán Vilas (Universidade da Coruña, besavi@mail2.udc.es)

1 Introducción

El proyecto "Base de datos léxica-semántica: unidades léxicas descriptivas y no descriptivas"(PGIDT99PXI10401B) tiene como uno de los objetivos principales la creación de una base de datos cuya información fundamental se centra en la combinatoria léxica de los lemas en español. A diferencia de otros investigadores en Lexicografía computacional que se han concentrado en la adquisición del conocimiento léxico a partir de diccionarios legibles por el ordenador, nosotros hemos optado por un léxico de nueva planta. Esta decisión se debe en gran medida a que los diccionarios actuales del español son todavía pobres en lo que se refiere a las relaciones léxico-semánticas sintagmáticas o *colocaciones*. El interés en el aspecto sintagmático es una de las características que nos distinguen de otros proyectos de Lexicografía computacional que se han centrado principalmente en atender a las relaciones semánticas paradigmáticas como la sinonimia, la antonimia y la meronimia, principalmente o en tratar aspectos más sintácticos como son la subcategorización o el régimen de los lemas, particularmente, verbales.

2 Noción de colocación y de función léxica

El término de *colocación* no es interpretado del mismo modo por los distintos investigadores: bien como combinación frecuente de palabras, en términos estadísticos, bien como una combinación en la que una palabra exige la presencia de otra para expresar un sentido dado, en términos más lexicográficos. Es esta interpretación lexicográfica la que utilizamos en el desarrollo de nuestra base de datos cuyo enfoque teórico subyacente es el de la Teoría Sentido-Texto (Mel'čuk y Polguère 1987) y más particularmente, la lexicología explicativa y combinatoria (Mel'čuk *et al*. 1995).

La herramienta lexicográfica de las funciones léxicas (FFLL) ha sido ampliamente rodada en los diccionarios explicativos y combinatorios del francés (DEC, vid. Mel'čuk *et al*. 1999). En cada artículo lexicográfico del DEC, se incluye una zona de coocurrencia léxica en donde se describen por medio de las FFLL las colocaciones en las que participa el lema como palabra llave de la FL. *Grosso modo*, una FL es una función **f** que asocia a una unidad léxica L_1 un conjunto de unidades léxicas cuasi-sinónimas $\{L_2\}$ que son escogidas en función de L_1 para expresar el sentido correspondiente a la FL **f**.

En términos de FFLL, una colocación L_1 (= base) y L_2 (= colocativo) se presenta como $\mathbf{f}(L_1) = L_2$, en donde L_1 es la PALABRA LLAVE de la correspondiente FL y L_2, su VALOR. A modo de ilustración, ofreceré ejemplos de algunas FFLL que describen colocaciones:

Magn(*lucha*) = *encarnizada*
Magn(*paciencia*) = *infinita*
Oper_1(*favor*) = *hacer* [*un* ~]
Oper_1(*paseo*) = *dar* [*un* ~]

Sus aplicaciones en el campo de la lingüística computacional también han dado sus frutos (vid. Fontenelle 1997, Heylen *et al*. 1994, Wanner 1996). Nuestro proyecto engarza con esta línea de investigación sobre la implementación de las FFLL y se inspira ampliamente en el proyecto canadiense DiCo, desarrollado por el equipo del Observatorio de Lingüística Sentido-Texto (vid. Polguère en prensa).

Creemos que la información sobre las colocaciones en que entra una unidad léxica dada es especialmente útil para los sistemas de traducción automática. Si el sistema reconoce una expresión como *dar un paseo* como una colocación, la traducción no podrá hacerse palabra a palabra. El sistema deberá comenzar por formatear la expresión lingüística a una fórmula funcional, como Oper_1(*paseo*) = *dar*. La traducción deberá hacerse en dos pasos:

1) traducción del argumento de la función:
paseo ==> *walk*;
2) búsqueda en el diccionario monolingüe del inglés de la función correspondiente:
Oper_1(*walk*) = *to take*

3 Estructura del artículo lexicográfico

En nuestra base de datos cada unidad léxica recibe un artículo lexicográfico completo. No existen, por tanto, lemas polisémicos. Si una palabra tiene varias acepciones, cada una de ellas corresponde a una unidad léxica distinta, si bien vinculada a las otras unidades léxicas que configuran el vocablo. Los lemas se limitarán a las bases de las colocaciones. De esta manera, no se

encontrará en nuestra nomenclatura el verbo *dar*, pero sí el nombre *pena* que es la base de la colocación *dar pena*.

Todo artículo lexicográfico dispone de tres zonas o secciones principales:

1) zona semántica, constituida por una etiqueta semántica que representa el significado central de la unidad léxica en cuestión (p. ej. 'sentimiento', 'hecho', 'acción', etc.) y una forma proposicional que indica la estructura de argumentos de la unidad léxica en cuestión ('pena de individuo X por hecho Y');

2) zona sintáctica, en donde se indica la realización superficial de los argumentos; p. ej. *la pena del niño <su pena>; la pena por <a causa de> la pérdida de su madre; pena de perder a su madre; pena por haber perdido a su madre*;

y 3) zona de relaciones léxico-semánticas, que pasamos a exponer con más detalle a continuación.

En dicha zona, se consignan todas las unidades léxicas que forman colocaciones junto con el lema. Las FFLL describen los colocativos aportando tanto información semántica como sintáctica. Según sea la clase de palabras del valor de la FL, podemos clasificarlas en:

1) FFLL adjetivales o adverbiales (Magn, Epit, Bon, Ver):

HONDA pena, *AMARGO desengaño*, *SINCERO agradecimiento*, *CÁLIDO agradecimiento*, etc.

2) FFLL verbales (Oper, Func, Labor, Real, Fact, Labreal, Caus, Liqu, etc.):

SENTIR pena por Y, la alegría REINA en Y, el remordimiento CORROE a X, *COSECHAR disgustos*, *DEPARAR muchos disgustos*, etc.

3) FFLL nominales (Sing, Mult, Figur, etc.)

MUESTRA de cariño, *MOTIVO de su pena*, *DOSIS de rencor*, etc.

4) FFLL preposicionales (Adv_i, Propt, Loc, etc.)

PARA alegría de todos, *llorar DE agradecimiento*, *POR miedo*, etc.

4 Estado del proyecto y planes futuros

Hasta el momento hemos elaborado alrededor de cincuenta entradas de vocablos del campo semántico de los nombres de emoción. Como media, los vocablos agrupan al menos dos unidades léxicas, pero hay vocablos como *pena* que incluyen hasta nueve. A pesar de la pequeña nomenclatura actual, la cantidad de unidades léxicas mencionadas como valores de FFLL es mucho mayor. A modo de ilustración, con la introducción de 24 vocablos, pueden encontrarse alrededor de 800 unidades léxicas relacionadas con los lemas. Contaremos con un mayor crecimiento de nuestro léxico en cuanto empecemos a adquirir semi-automáticamente colocaciones, a partir del análisis de los corpus. Una tarea en desarrollo es la formulación de criterios heurísticos que faciliten la asignación automática de una FL a una colocación dada.

Con todo, la información ya disponible en nuestra base de datos[1] nos permite realizar consultas que pongan en relación etiquetas semánticas con colocaciones, así como hacer consultas a la base de datos en sentido inverso, del colocativo a la base, de modo que el sistema nos devuelva cuáles son todos los nombres que se combinan con un verbo o un adjetivo dado para formar una colocación.

Referencias

Fontenelle, T.(1997): *Turning a Bilingual Dictionary into a Lexical-Semantic Database*, Tübingen: Niemeyer

Heylen, D., K.G. Maxwell y M. Verhagen (1994): "Lexical Functions and Machine Translation", COLING, Japón, pp. 1240-1244.

Mel' uk, I., N. Arbatchewsky-Jumarie, L. Iordanskaja, S. Mantha y A. Polguère (1999): *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*, Montréal: Les Presses de l'Université de Montréal.

Mel' uk, I., A. Clas y A. Polguère (1995): *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve: Duculot.

Mel' uk, I. y A. Polguère (1987): "A Formal Lexicon in Meaning-Text Theory (or How to Do Lexica with Words)", *Computational Linguistics*, 13, 3, pp. 276-289.

Polguère, A. (en prensa): "Toward a Theoretically-Motivated General Public Dictionary of Semantic Derivations and Collocations for French", *Proceedings of 9th Euralex*.

Wanner, L. (ed.) (1996): *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam, Philadelphia: John Benjamins.

---

[1] Agradecemos especialmente el trabajo informático realizado por Oscar Sacristán quien fue el encargado de diseñar la base de datos en el programa Access 2000.