Explotación informática de una base de datos multilingüe de unidades fraseológicas[1]

Pedro MOGORRÓN HUERTA
Universidad de Alicante
pedro.mogorron@ua.es

Resumen: En este trabajo presentamos los diferentes pasos que el grupo de investigación FRASYTRAM está llevando a cabo para la elaboración de una base de datos multilingüe sobre las unidades fraseológicas, concretamente las construcciones verbales fijas, mediante la consulta de fuentes no sólo escritas (obras lexicográficas y fraseográficas) sino también orales. Mostramos las posibilidades que ofrecen las tecnologías de la información y de la comunicación para el tratamiento de la base de datos. La informatización de las TIC está permitiendo una verdadera revolución en el tratamiento, la selección, la enseñanza y la traducción de estas formas, gracias a numerosos programas y aplicaciones específicos que permiten sistematizar su clasificación y facilitar tanto su uso como su difusión. Esta base de datos es de gran utilidad tanto para la traducción como para la enseñanza de lenguas, pues contiene las construcciones verbales más empleada y, por tanto, las que pueden constituir el mínimo fraseológico.
Palabras clave: Fraseología. Enseñanza de lenguas. Traducción. TIC.

Titre : « Explotation informatique d'une base de données multilingue d'unités phraséologiques ».
Résumé : Dans ce texte nous présentons les différents travaux que le groupe de recherche FRASYTRAM est en train de réaliser afin d'élaborer une base de données multilingue d'unités phraséologiques, plus précisément des expressions verbales fixes, en consultant des sources non seulement écrites (des ouvrages lexicographiques et phraséographiques) mais aussi orales. Nous montrons les possibilités qu'offrent les technologies de l'information et de la communication pour le traitement de la base de données. L'informatisation des TICE est en train de permettre une véritable révolution dans le traitement, la sélection, l'enseignement des langues ainsi que dans la traduction de ces formes, grâce aux nombreux programmes informatiques et aux applications spécifiques qui permettent de systématiser leur classification et de faciliter leur usage et leur diffusion.
Mots-clés : Phraséologie. Enseignement de langues Traduction. TICE.

[1] El siguiente trabajo se inscribe en el marco del proyecto de investigación FFI2011-24310 «Estudio lingüístico, diatópico y traductológico de las construcciones verbales fijas más usuales en español».

Title: «Informatic explotation of a multilingual database of idioms».
Abstract: This work describes the different steps taken by the research group FRASYTRAM in the elaboration of a multilingual database of idioms, specifically fixed verbal expressions by means of not only written (lexicographical phraseographical works) but oral sources. The aim is to show the potential of information and communication technologies and their application to databases. Computers and text analysis software applied to ICT will mean a revolution in the processing, selection, teaching and translation of these forms, since they allow systematization and classification, and facilitate their use and difussion. This database is of great use for translation and for teaching of languages, because it has the most frequently used verbal constructions, and therefore those that could belong to the paremiological minimum.
Keywords: Phraseology. Learning languages. Translation. ICT.

INTRODUCCIÓN

Todas las lenguas conocidas vienen determinadas, no solamente por las reglas de libre composición que establecen qué elementos léxicos pueden combinarse entre sí, sino también, por un gran número de unidades fraseológicas (UF) utilizadas frecuentemente en todos los actos de comunicación y caracterizadas por la fijación[2]. Así, en español, para expresar que una persona habla mucho, los usuarios podrán decir: a) Juan habla mucho; b) Juan es muy parlanchín; c) Juan no para de hablar; d) *Juan habla como un loro* (DUE); e) *Juan habla por los codos* (DUE); f) *Juan habla por siete* (DRAE); g) *Juan echa un perico* (AoMex), h) *Juan cacha la víbora* (DHDA), etc.

Los ejemplos d, e, f corresponden a UF que forman parte de un fondo cultural y lingüístico común. Los ejemplos g y h son variantes diastráticas que corresponden a variantes mexicanas y argentinas respectivamente y ya no forman parte de este fondo cultural y lingüístico común en el que las UF se han ido adquiriendo y almacenando dentro de la memoria cultural y colectiva de la sociedad y a su vez dentro de la memoria individual de los usuarios. Pero, por otro lado, es innegable que pertenecen al español. Será, pues, necesario que el interlocutor y el locutor posean simultáneamente ese fondo común que reactivan e individualizan en cada acto de comunicación. En caso contrario, el usuario de la lengua se encontrará con grupos de palabras que no podrá comprender por su sentido idiomático y deberá recurrir a la ayuda de los diccionarios para intentar descifrar esa posible UF.

Además conviene señalar y destacar que las UF se utilizan con un objetivo discursivo muy claro. Representan el deseo del locutor de utilizar en el acto de comunicación una determinada fórmula refrendada por la mayor parte de la comunidad lingüística, sumándole de esa forma un matiz cultural, social, generacional, etc., en vez de utilizar un término neutro o una perífrasis verbal.

[2] Sobre las características de las UF, véanse, entre otros, los trabajos de Corpas (1996), Ruiz (1997), García-Page (2008).

Desde nuestra docencia en Traducción, hemos podido comprobar reiteradamente como muchas UF españolas eran desconocidas para muchos usuarios, por lo que suponían un verdadero reto tanto de comprensión como traductológico. Tras numerosas consultas de diccionarios para averiguar el significado de UF así como su(s) equivalente(s) en otro(s) idioma(s), hemos podido observar la enorme cantidad de UF que no están presentes en los diccionarios monolingües y bilingües más conocidos. Por ejemplo:

- *andar a grillos* (ocuparse en cosas inútiles o baladíes); *andar con la barba por el suelo* (ser muy anciano o estar decrépito) vienen registradas en el DRAE, pero no figuran en el DUE ni en el LARBI;
- *arder en fiestas* (estar una ciudad muy animada con la celebración de festejos) y *echar los pies por alto* (descomponerse o enfurecerse*)* aparecen en el DUE, pero no en el DRAE ni en el LARBI
- *dar con el codo* (hacer a alguien una seña de esa manera) figura en el DUE y en el LARBI, pero no en el DRAE, que incorpora en cambio las variantes paradigmáticas *dar de codo* o *dar del codo*.
- *estar a mesa y a mantel* (comer diariamente con él y a su costa) viene registrada en el DRAE y en el LARBI (*se faire nourrir par qq'un*, *vivre aux frais de qq'un*), pero no en el DUE que incluye la variante *tener a alguien a mesa y a mantel* (darle de comer gratis a diario).

Del mismo modo hemos podido comprobar la existencia de diccionarios fraseológicos monolingües que, pese a caracterizarse por su falta de rigor y exhaustividad, suelen incorporar UF pertenecientes a diferentes categorías fraseológicas.

La gran cantidad de UF que no aparecen en los diccionarios tanto monolingües como bilingües nos ha llevado a pensar en la necesidad y utilidad de elaborar un diccionario fraseológico que permitiese consultar sino todas, al menos una gran mayoría de las UF presentes en nuestro idioma para poder permitir a los usuarios encontrarlas y comprenderlas.

1. DELIMITACIÓN TEÓRICA DE LA BASE DE DATOS.

Debido a la gran variedad de tipos de UF presentes en las lenguas[3] y ante la imposibilidad de poder tratar todas estas categorías fraseológicas, decidimos limitar nuestro estudio a las construcciones que venimos llamando «construcciones verbales fijas» CVF (Mogorrón, 2008 y 2010) y en las que hemos integrado las siguientes construcciones fijas:

[3] Colocaciones verbales, nominales, construcciones con verbos soporte, locuciones (verbales, adjetivas, nominales, adverbiales, conjuntivas), enunciados fraseológicos, paremias, fórmulas rutinarias o pragmatemas, etc. (Sevilla Muñoz, 1993; Corpas Pastor, 1996; Ruiz Gurillo, 1997; García-Page 2008).

1) locuciones verbales, *hacerse el sueco* (DUE); *afeitar un huevo en el aire* (DTDFH);
2) colocaciones verbales[4], *guiñar un ojo* (DUE); *derramar lágrimas*; *formular una pregunta*, (DUE);
3) verbos soportes[5], *dar un paseo* (DUE);
4) construcciones verbales comparativas, *dormir como un tronco* (DUE); *llorar como una Magdalena* (DUE).

Somos conscientes de que estos tipos de construcciones que estamos enumerando no pertenecen para muchos teóricos de la lengua a la misma categoría de UF (Gross, 1996: 69-88; Corpas, 1996; Ruiz Gurillo, 1997; García–Page, 2008: 12-22; Zuluaga, 1980), y que presentan diferencias estructurales y propiedades diferentes como la idiomaticidad, la composicionalidad[6]; sin embargo, tienen muchos puntos en común: todas ellas son pluriverbales (se componen de al menos dos unidades léxicas) y están compuestas por un verbo más un complemento y un cierto grado de fijación lingüística. En efecto, no debemos olvidar que las UF son ante todo *complejos sintagmáticos fijos* (Ruiz Gurillo, 1997: 104).

2. ELABORACIÓN DE LA BASE DE DATOS.

Una vez definidas las construcciones que iban a componer nuestra base de datos (BD), el grupo de investigación FRASYTRAM de la Universidad de Alicante está elaborando desde 2005 una base de datos multilingüe de construcciones verbales fijas (CVF)[7]. El primer paso consiste en elaborar un diccionario electrónico, lo más exhaustivo posible, en forma de base de datos, de las CVF españolas señaladas. La construcción de esta base de datos de CVF, se está llevando a cabo mediante la consulta de un numeroso grupo de diccionarios monolingües y bilingües (véanse en las referencias bibiográficas) y la incorporación de las creaciones de la lengua activa no encontradas en los diccionarios consultados pero de uso habitual en la lengua española. En este caso, recurrimos a la competencia fraseológica de los miembros del grupo de investigación a la par que realizamos numerosas consultas en internet y en corpus textuales de la Real Academia

[4] Las colocaciones verbales son combinaciones sintagmáticas en las cuales se establece una relación de solidaridad léxica entre sus componentes.
[5] Las «construcciones con verbos soporte» son construcciones en las que el verbo no tiene significado y la función de predicado viene desempeñada por el sustantivo.
[6] En el caso de las construcciones comparativas que algunos llaman elativas y, en algunos casos adverbiales, se trata de una construcción que se usa casi siempre con algún verbo estableciendo una asociación preferencial, por lo que también las recopilamos para permitir al usuario saber qué verbo es el que se utiliza preferencialmente con esas construcciones.
[7] http://labidiomas3.ua.es/phraseology/login/login.php

Española (CREA/CORDE[8]). Por ejemplo: *echar fuego por la boca*; *echar las* [*papa*s, *las peras*]; *pasar más hambre que un maestro de escuela; pasárselo pipa; ponerle a alguien el culo como un tomate; comerse los mocos* (pasar escasez de cualquier tipo)[9].

Esas consultas nos han permitido incluir, hasta la fecha, en la base de datos unas veinte mil CVF. Para la consulta de los diccionarios utilizados, seguimos el siguiente orden por ir en proporción con el número de UF recopiladas: DUE, DRAE, DTDFH, DFDEA, DEA y el EPM. Después los restantes diccionarios indistintamente. Cabe destacar que cerca de once mil CVF aparecen en uno de los dos diccionarios más usuales del español (DUE y DRAE) [10], lo que significa que el 46% de las CVF no aparecen en el DUE y en el DRAE y que el 11,7% de las CVF no figura en ninguno de los diccionarios consultados.

Las versiones comerciales de algunos de estos diccionarios en soporte informático, no ha supuesto ningún cambio apreciable en la estructura o en el contenido de estas obras, pues las únicas diferencias que se han podido observar han sido la rapidez de las consultas a realizar, la posibilidad de navegar por los artículos así como la de seleccionar y buscar estructuras con pequeños localizadores, ya que tanto la estructuración de la información como la de los contenidos siguen siendo prácticamente similares.

Cabe mencionar, por otro lado, la reciente aparición en el mundo de la lingüística de otro tipo de diccionarios cuya elaboración está relacionada con posibles aplicaciones al tratamiento automático de textos en el campo de la lingüística aplicada. Se trata de los diccionarios electrónicos. Al contrario de lo que había ocurrido con la elaboración y el contenido parcial de los diccionarios generales clásicos, ya desde sus inicios, la concepción de este nuevo tipo de diccionarios se está realizando asumiendo como objetivo a alcanzar, el de la máxima exhaustividad posible. En efecto, un diccionario electrónico no es una sencilla lista alfabética muy completa de palabras simples o compuestas que se podrá posteriormente aplicar al análisis automático de textos. No. Un diccionario electrónico es mucho más que eso. Se trata de una base de datos que contiene una enorme cantidad de datos recogidos por los lexicógrafos que deberá ser gestionada por la informática en función de las necesidades que se hayan programado y que se deseen alcanzar.

[8] http://corpus.rae.es/creanet.html y http://corpus.rae.es/cordenet.html, respectivamente.

[9] Existe un desfase considerable entre el continuo e imparable proceso de remodelación de los actos de comunicación constantemente obligados a readaptarse y el material que los lexicógrafos insertan en los diccionarios. En efecto, por un lado están las UF de la lengua clásica, con frecuencia ya en desuso. Por otro lado, se halla el uso presente, marcadamente innovador de la lengua, que se debe ir transformando y adaptando a los constantes cambios que se producen en la lengua por los usos generacionales y las diferentes necesidades y realidades sociales que se plasma en nuevos vocablos y expresiones.

[10] Lo que queremos dejar patente aquí es la presencia/ausencia de estas formas en los que se consideran los dos mejores diccionarios del español actual.

Para poder entender las posibilidades de los diccionarios electrónicos hay que saber que se basan en torno a dos piedras angulares. Por un lado tendremos una base lingüística y por otro lado la parte informática que se va a encargar de gestionar lo más eficazmente posible toda la cantidad de información lingüística recogida por los especialistas en lengua. Ésta podrá ser por lo tanto: morfológica, sintáctica, fonética, semántica, etc. Sin embargo, la parte informática solamente podrá analizar la información que la base de datos lingüística haya recopilado y estructurado y realizar procesos selectivos de búsqueda sobre ésta. El lingüista debe saber lo que necesita el informático, y el informático a su vez lo que desea conseguir el lingüista. Una vez estén definidos entre los dos, los objetivos a alcanzar, se deben crear una serie de ficheros informáticos que contendrán toda la información en forma de entradas léxicas (simples o complejas), en forma de códigos morfológicos identificables por los programas. Las aplicaciones contempladas por los diccionarios electrónicos son muy numerosas, pero van a depender en gran parte de la información y del tratamiento que se le haya dado a ésta.

A continuación mostraremos la elaboración de la base de datos de CVF que estamos realizando en formato Excel con la inclusión de numerosas informaciones de carácter léxico, sintáctico, semántico, cultural, etc. que permitirán posteriormente con la ayuda de los filtros de la aplicación realizar búsquedas y selecciones múltiples relacionadas con los temas introducidos y que puedan interesar al investigador, al lingüista y al usuario.

2.1. Información lexicográfica.

En esta fase de la elaboración de la base de datos anotamos en columnas diferentes los verbos y los sustantivos. De esta forma lanzando una búsqueda con los filtros podemos pedirle a la BD que nos seleccione todas aquellas CVF en las aparece un determinado sustantivo o verbo. Así, la búsqueda de CVF en las que aparezca la palabra «corazón» nos permite observar que más de un centenar de UF en la BD contienen esta palabra.

2.1.1. La variación.

La elaboración de este tipo de BD nos permite afirmar que existen dos tipos de CVF:

- Las que no permiten variación alguna de los elementos léxicos que las componen: *enterrar el hacha de guerra* (DUE)*; quemar las naves* (DUE); *hacer alguien de tripas corazón* (DUE); *írsele a alguien el santo al cielo* (DUE)*, ladrar a la luna* (DUE), *liarse la manta a la cabeza* (DUE).
- Aquellas CVFS que presentan variaciones léxicas o paradigmáticas de algunos de sus componentes: [estar, ir, ponerse] *de veinticinco alfileres* (DUE)*;* [andar(se), echar, irse, marcharse, salir] *por los cerros de Úbeda* (RAE); [echar, lanzar] *las campanas al vuelo* (DUE); *llamarse [a andana* (RAE), *a antana* (DUE), *andana*

(DUE), *antana* (DUE)]; *meter [el hocico, la nariz, las narices, los hocicos] en algo*.

En efecto, hemos podido comprobar que existen efectivamente numerosas CVF en las se aprecian variantes que pueden intercambiarse sin que el significado de estas construcciones varíe. Ej.: *abrirse [paso, camino] a codazos* (DT); *buscarle [cinco pies, tres pies] al gato* (DUE); *tirar a ventana [conocida, señalada]* (DRAE); etc. Además, tal y como podemos observar en el cuadro 1, los diferentes diccionarios pueden también presentar variantes diferentes. No todos los diccionarios incluyen variantes paradigmáticas, o todas las variantes paradigmáticas, con lo cual, si no se hace una búsqueda exhaustiva, se puede consultar un diccionario que no incluya variante(s) y pensar que esa construcción no tiene variantes cuando otro u otros diccionarios las presentan. Así, para la CVF *coger el toro por los cuernos* encontramos en los diccionarios consultados las siguientes formas:

DUE	[agarrar / coger] el toro por los cuernos
RAE	coger al toro por [las astas / los cuernos]
EPM	[agarrar / coger / tomar] el toro por los cuernos
LARBI	[agarrar / coger / tomar] el toro por los cuernos
DT	agarrar al toro por los cuernos

Cuadro nº1

Hemos reflejado esas variantes de la forma siguiente en una columna de la BD que indica las posibles variantes de las CVF: [*agarrar*, *coger*, *tomar*] *el toro por los cuernos*, [agarrar*] al toro por* [*los cuernos*], *coger al toro por las astas*. Además queremos señalar que también hemos encontrado en internet *tomar el toro por las as*tas y [*agarrar*, *coger*, *tomar*] *el toro por los cachos*, sobre todo con páginas webs de Hispanoamérica[11].

Hemos optado por reflejar todas y cada una de las posibles variantes recogidas en los diccionarios en la BD. Para ello, cada variante figura como una entrada individual indicando la fuente documentada en la que la hemos encontrado:

agarrar al toro por los cuernos (DT);
agarrar el toro por los cachos (internet, Hispanoamérica)
agarrar el toro por los cuernos (DUE);
coger al toro por las astas (DRAE);
coger al toro por los cuernos (DRAE);
coger el toro por los cachos (internet, Hispanoamérica)
coger el toro por los cuernos (DUE);
tomar el toro por las astas (internet)
tomar el toro por los cachos (internet, Hispanoamérica)
tomar el cuerno por los cuernos (EPM);

[11] Esta variante aparece en internet en más de 40.000 entradas.

Para corroborar la afirmación de que las variantes son una producción lingüística frecuente, nos apoyaremos en las cifras de nuestra base de datos. Las cifras de las que disponemos actualmente, nos indican que el 53% de las CV permiten una o varias variantes de sus componentes[12]. Se trata, pues, de un fenómeno mucho más importante de lo que parecía y de gran importancia de cara a la utilización de estos datos en programas de TAL y de Traducción automática. Esta información queda reflejada con un desarrollo de las posibles variantes y con unas siglas que indican en cada caso si se trata de una variante del sustantivo, del verbo, de los modificadores, ortográfica, etc.. Para ello, en dos columnas de la BD indicamos respectivamente cada una de las posibles variantes encontradas hasta la fecha así como las siglas que le corresponden.

2.1.2. La categoría fraseológica

Otra de las columnas de la BD, clasifica cada una de las CVF en locución, construcción con verbo soporte o colocación, etc. Se trata de una información valiosa para los fraseólogos y que deseamos desarrollar en el marco del proyecto de investigación FFI2011-24310 que estamos desarrollando.

2.1.3. La polisemia.

Numerosas UF presentan un fenómeno de diversificación del significado. La elaboración de la BD a partir de las CVF recopiladas en numerosos diccionarios nos ha permitido encontrar numerosas expresiones polisémicas con idéntica forma pero con diferente significado. Se trata de un fenómeno que tampoco ha sido tratado con exhaustividad en los diccionarios en soporte papel y en los diccionarios electrónicos. Así para la expresión *aguzar los dientes* hemos encontrado los siguientes significados en los siguientes diccionarios monolingües, bilingües y fraseológicos:

- DRAE: «disponerse para comer, cuando está pronta e inmediata la comida».
- DUE, DFDEA: no aparece.
- GDFH de Larousse: «significa prepararse para comer, cuando está lista la comida».
- En la EPM hallamos estas definiciones: 1) «prepararse para la comida»; 2) «ansiar una cosa»; 3) «apropiarse indebidamente de una cosa que se administra o custodia»; 4) «murmurar, refunfuñar»; 5) «enfrentarse a las dificultades de un asunto»; 6) «criticar a alguien».
- En el LARBI encontramos un significado nuevo: «*aguzarse los dientes = se faire la main*», es decir según el AR: «s'exercer à un travail réclamant de l'habileté manuelle».

[12] Pensamos que la búsqueda de estas UF en bases textuales de gran tamaño permitirá encontrar con toda seguridad más variantes que no figuren en los diccionarios consultados, pero que sean muy usuales en la lengua.

La polisemia produce en este caso un factor de opacidad. En el apartado de expresiones polisémicas conviene también señalar la opacidad en expresiones diatópicas que son utilizadas en alguno de los países de habla hispana. Por ejemplo la expresión *doblar la esquina* aparece en los diccionarios consultados con los siguientes significados:

doblar la esquina	morirse	LARBI
doblar la esquina	girar de una calle a otra	MM
doblar la esquina	desaparecer	MM
doblar la esquina	cambiar de tema, pasar a tratar un asunto o tema diferente.	DTDFH (Cuba)

Cuadro nº 2

Si como hemos podido apreciar, el usuario no conoce muchas de las UF, resulta obvio deducir que, tampoco conocerá muchos de los significados de estas expresiones polisémicas que a su vez tampoco aparecen en numerosos diccionarios que no tratan la polisemia en profundidad. Hemos catalogado hasta la actualidad más de 1600 expresiones polisémicas con unas 4300 acepciones. Se trata pues de un fenómeno mucho más frecuente de lo que hubiera podido parecer en un principio y que puede sin lugar a dudas plantear numerosos problemas de interpretación y de uso en los programas de traducción automática ya que el tema de la polisemia y de la interpretación correcta de los componentes; del significado y de la posible doble lectura con la ambigüedad que conlleva son también temas recurrentes en lingüística informática y computacional así como en la traducción automática (TAO) debido a los numerosos problemas que plantean.

1.1. Información sintáctica.

Las aplicaciones contempladas por los diccionarios electrónicos son muy numerosas, pero van a depender en gran parte de la información y del tratamiento que se le haya dado a ésta. En efecto, la utilización de la información en tratamiento automático de textos exige que las palabras pertenecientes a los textos que se han introducido, estén etiquetadas de manera que el diccionario pueda catalogarlas. A partir de técnicas basadas en el procesamiento del lenguaje natural, se han desarrollado sistemas para la lematización, es decir para el etiquetado automático morfológico y sintáctico de los textos de un corpus, que consisten en la lectura y en la división del texto en unidades relevantes que serán más tarde utilizadas para trabajar el análisis de la palabra. Las características que normalmente se indican durante este etiquetado hacen referencia a:

- aspectos de estructuras de texto: marcas tipográficas, divisiones textuales, párrafos, citas, títulos.
- las propiedades morfosintácticas de la palabra.

- funciones sintácticas de cada constituyente y representación por medio de árboles sintácticos, etc.

Si observamos ahora la base de datos, la novedad en este caso aparece en las columnas C y D en las que respectivamente aparece para cada CVF la estructura sintáctica con las normas del lexique-grammaire de Maurice Gross (1996) y el nombre de la clase que corresponde a esa estructura. Por ejemplo:

abandonar el barco	N0 V Ddef C1	C1D
abandonar el campo	N0 V Ddef C1	C1D
abandonar el campo	N0 V Ddef C1	C1D
abandonar el lecho	N0 V Ddef C1	C1D
abrirse un abismo entre	N0 V dindef C1 Prep N	C1IPN
abrirse camino a codazos	N0 V C1 Prep C2	C1P2

Cuadro nº 3

Estas estructuras sintácticas facilitarán posteriormente la elaboración de arboles sintácticos que se usarán en programas de tratamiento automático del lenguaje.

1.2. Información semántica

La consulta de los numerosos diccionarios analizados para la elaboración de la base de datos y la búsqueda de equivalentes fraseológicos nos ha permitido detectar que existen frecuentemente CVF parasinónimas. Estas CVF parasinónimas forman grupos heterogéneos imprevisibles en cuanto al número de integrantes que pueden ir desde un par de UF, hasta varias decenas de expresiones. Así, para decir que una persona es insensible hemos encontrado: *ser de bronce* (DUE), *no tener corazón* (DUE), *tener el corazón de piedra* (EPM).

Con el significado de estar muy delgado, la BD contiene unas sesenta expresiones, de las que reproducimos a continuación algunos ejemplos: *estar chupado* (EPM); *estar como un fideo* (DFDEA); *estar como una espátula* (EPM); *estar delgado como un palo* (DFDEA); *estar hecho una momia* (DUE); *ser un palillo* (DFDEA). En la BD hemos utilizado la «misma definición» para cada una de estas UF.

La aparición de tantos parasinónimos nos hizo reflexionar, como docente de traducción, al ver la dificultad de seleccionar uno de ellos como equivalente para una CVF en otro idioma en el que posiblemente para muchos conceptos, actos, descripciones muy usuales existirían también numerosas representaciones léxicas y UF. Estas reflexiones nos han llevado a plantear la necesidad de elaborar una herramienta que fuese de utilidad para los traductores. En efecto, la traducción de estas formas no ha sido tratada en profundidad, hasta ahora, por la lexicografía bilingüe. El procedimiento tradicional utilizado para reproducir en otra lengua una UF, ha consistido o bien en utilizar un diccionario bilingüe para transcribir, si con un poco de suerte la forma viene tratada en el

diccionario, o bien si el traductor o el usuario posee una buena competencia fraseológica, en poner ésta última a prueba para encontrar una forma más o menos equivalente en la otra lengua. Esas dificultades repetimos, nos han llevado a plantear la necesidad de realizar una clasificación onomasiológica de las CVF. En efecto numerosos usuarios e investigadores pueden estar interesados en obtener información acerca de todas las UF que pertenezcan a un mismo campo semántico, con los parasinónimos, los antónimos. Esto conlleva realizar una clasificación semántica para cada CVF.

El siguiente paso en traducción implica buscar los equivalentes de traducción para todas estas CVF. Para ello, estamos elaborando una aplicación informática en la que la información onomasiológica y semántica de los cuadros anteriores aparece de manera muy intuitiva y permite al usuario realizar las búsquedas rápidamente. La aplicación presenta unos campos semánticos muy amplios: carácter-forma de ser; comunicación, climatología, deporte, descripción física.

Cada uno de estos campos semánticos se divide en subcampos semánticos. Así las CVF que pertenecen a la descripción física se agrupan en aspecto, belleza, fealdad, color de la piel, complexión,…

Esta aplicación nos permitirá encontrar una CVF a partir de sus componentes, de su pertenencia a un campo semántico, de su definición, de palabras clave. Una vez seleccionado el modo de búsqueda, y encontrada la CVF que nos interese, encontraremos también todas las CVF parasinónimas.

Para cada UF, los usuarios podrán también consultar una serie de informaciones: fuente en la que se ha recopilado, frecuencia de uso, valor diatópico, contextos, nivel de lengua, etc. que le permitirán encontrar una UF equivalente en función de los valores de la expresión a traducir Para buscar los equivalentes bastará con pulsar las pestañas de los idiomas señalados en la parte superior de la aplicación para que automáticamente aparezcan en el idioma requerido.

1.3. Nuevas aplicaciones en la Base de Datos.

Dentro del proyecto de investigación que estamos llevando a cabo, (ver nota a pie de página nº 1), deseamos dar un salto cualitativo en la investigación que venimos desarrollando y realizar una investigación innovadora que permita transformar esta bases de datos de expresiones en una potente herramienta polivalente y versátil. Para ello vamos a incrementar el número de CVF incluyendo variantes diatópicas de Argentina, México, Colombia y Perú)[13]. Reproducimos a continuación un ejemplo de la BD de expresiones de origen argentino.

[13] En el caso de lenguas como el español, el inglés, el francés, el portugués que son lenguas (co-)oficiales en numerosos países, pensamos que es de gran importancia incluir las producciones de

- seleccionar las 1500 CVF más frecuentes, tanto en español como en las variantes diatópicas señaladas, apoyándonos gracias a la lingüística informática en una base textual y en los buscadores de internet. Se trata sin lugar a dudas de la parte más innovadora de nuestra investigación. Para verificar el conocimiento fraseológico de los alumnos, tanto españoles como extranjeros, así como la validez del material didáctico que vamos a elaborar, hemos recibido el acuerdo de las Universidades de: Paris 13; Bari; Goettingen; Napoli (Suor Orsola Benincasa); Benemérita Universidad Autónoma de Puebla en México; Antioquía (en Colombia).
- disponer de ejemplos contextualizados que permitan ver materializados el uso de palabras o expresiones dada la importancia que tiene para la enseñanza de la lengua y para los usuarios o traductores. Por ello:
  a. El LDI de Paris 13 con el que cooperamos estrechamente y que participa en el proyecto I + D + I ha creado una herramienta informática de recopilación de textos mediante sindicación de contenidos (RSS) de periódicos digitales, que permite recuperar automáticamente y regularmente los documentos textuales de los periódicos, así como almacenarlos en una base de datos textuales centralizada.

     Como botón de muestra de la herramienta, reproducimos a continuación una muestra de la extracción del contenido del periódico *El Mundo* disponible en línea, que contiene los textos publicados el día 20 de septiembre de 2011, con el fin de obtener resultados sobre la lengua española de registro estándar lo más actual posible. A través de unas operaciones informáticas, el programa aspira automáticamente los textos de diferentes periódicos a través de la herramienta «RSS Corpus Builder». Una vez extraídos los textos y clasificados en secciones bastará con hacer clic sobre el archivo ejecutable de la aplicación para que se aspiren los archivos en formato txt en la carpeta «corpus». Esta operación se puede programar para que se inicie automáticamente cada día y poder disponer, en poco tiempo, de un corpus denso.
  b. Estamos recopilando y adquiriendo miles de obras literarias en español en formato txt para poder realizar la búsqueda de contextos a la vez en los periódicos y en las obras literarias. Clasificaremos estas obras en función de su tipología, época, país de origen.
- Uso de herramientas para detectar la frecuencia de uso de las CVF de la BD.

Los dos corpus textuales que se están elaborando aparecerán en formato txt. Nos van a permitir buscar las CVF más frecuentes que figuran en nuestra BD. La lingüística de corpus ha generado un amplio número de herramientas que permiten el análisis de textos. Por un lado podemos destacar las herramientas centradas en la construcción de sistemas de etiquetado y análisis morfosintáctico (ejemplo: http://igm.univ-

cada uno de los países. Ya contamos actualmente con más de 1000 CVF en el caso de Argentina y de México, y estamos procediendo a recopilar CVF de Perú y Colombia.

mlv.fr/~unitex/index.php?page=3#)[14]. Por otro lado existen numerosos programas informáticos procesadores de textos que se caracterizan por permitir el análisis de los textos desde la perspectiva de las frecuencias, agrupamientos y concordancias de unidades léxicas[15]; por ejemplo: Wordsmith tools (http://www.lexically.net/wordsmith/), y AntConc (http://www.antlab.sci.waseda.ac.jp/)[16].

El programa Unitex (de descarga gratuita y abierta al público) funciona a partir de textos etiquetados o de gráficos que se encargaran de detectar la presencia de las UF contenidas en el diccionario. Con las UF que figuran en el cuadro nº 4 se elaboran gráficos:

Verbo	**Expresión**	**Verbo**	**Expresión**
abajar	abajar el casco	**abrir**	abrir algo de par en par abrir (bien) los ojos abrir (nuevos) horizontes abrir (una puerta) de par en par abrir a alguien como a un cerdo abrir alguien los oídos abrir boca abrir brecha abrir calle abrir camino abrir consulta abrir las puertas (a algo, para que x) abrir las zanjas abrir(le) los brazos a abrir los oídos (a?) abrir los ojos abrir los ojos como platos abrir ojos como platos abrir plaza abrir puerta (a algo, para que x) abrir un crédito a abrir tanto ojo abrir unos ojos como platos abrir(le) camino (a x, para x?) abrir(le) cancha a alguien
abandonar	abandonar a alguien a pos suerte abandonar a alguien en manos de abandonar el barco abandonar el campo abandonar el lecho abandonar la lucha (armada, + adj activa, pasiva) abandonar la partida abandonar las armas abandonarse en brazos de		
abarcar	abarcar demasiado		
abatir	abatir banderas		
ablandar	ablandar las piedras ablandársele el alma a alguien		
abonar	abonar el terreno (para, a)		
aborrecer	aborrecer de muerte a alguien aborrecer los huevos		
abrasar	abrasar la sed abrasarse las pajarillas abrasarse los pájaros abrasarse vivo		
abrazar	abrazar una causa		
abrigar	abrigar dudas abrigar sospechas abrigar una sospecha		

[14] Las diferentes versiones estables se encuentran disponibles en la página de la Universidad Paris-Est Marne-la-Vallée: http://igm.univ-mlv.fr/~unitex/index.php?page=3#

[15] Para que los resultados puedan considerarse fiables, los corpus textuales deben tener unas grandes dimensiones (varios centenares de millones de palabras).

[16] *Antconc* y *Wordsmith* son programas para el análisis de concordancias, es decir, que permiten mostrar el contexto de aparición de palabras clave en un texto o conjunto de textos.

Verbo	Expresión	Verbo	Expresión
abrir	abrir cancha alguien abrir(le) el alma a otra persona abrir(le) el camino (a x, para x?) abrir(le) el corazón a otra persona abrir(le) la cabeza a alguien abrir(le) la puerta (a algo, para que x) abrir(le) las puertas a alguien abrir(le) los ojos a alguien abrir(le) pos alma a otra persona abrir(le) pos corazón a otra persona abrir(le) pos pecho a otra persona abrir(le) una puerta (a algo, para que x) abrir(se) paso abrir(se) un abismo entre no abrir la boca abrir(se) paso alguien abrir(se) un abismo entre abrir el apetito abrir el baile abrir el compás abrir el día abrir el ojo, (a, para, con el fin de) abrir el paraguas abrir el pico abrir el tiempo abrir en canal abrir fuego contra abrir fuego sobre abrir la boca abrir la caja de los truenos abrir la caja de pandora abrir la corona abrir la espita	**abrir**	abrir la gloria abrir la mano abrir la mano a algo / en algún tema abrir la marcha abrir la puerta (a algo, para que x) abrir las ganas abrir las puerta (a algo, para que x) abrirse de capa abrirse de piernas abrirse de piernas (ante?) abrirse la cabeza abrirse las venas abrirse paso abrirse paso a codazo limpio abrirse paso a codazos abrirse un abismo entre abrírsele a alguien las carnes abrírsele el cielo a alguien abrírsele la boca a alguien
		absolver	absolver a alguien a cautela absolver a culpa y pena
		abultar	abultar lo que un comino
		abundar	abundar como hongos en año de lluvias
		aburrir	aburrirse hasta no poder más aburrir a las ovejas aburrir hasta a las ovejas aburrirse como un hongo aburrirse como una mona aburrirse como una ostra aburrirse como una sota

Cuadro nº 4

El método consiste en realizar un gráfico por cada verbo incluido en los grupos de expresiones. Así, tendremos 15 gráficos, correspondientes a los verbos *abajar*, *abandonar*, *abarcar*, *abatir*, *ablandar*, *abonar*, *aborrecer*, *abrasar*, *abrazar*, *abrigar*, *abrir*, *absolver*, *abultar*, *abundar* y *aburrir*. Algunos de ellos, como por ejemplo en verbo «abajar», precisan de un gráfico simple, pues solo tienen una expresión asociada al verbo:

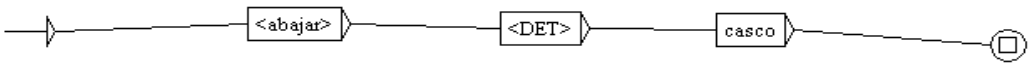

Gráfico nº 16.

Sin embargo, otros verbos como «abrir» poseen un gráfico bastante complejo, pues existen múltiples posibilidades de expresiones que contienen ese verbo ():

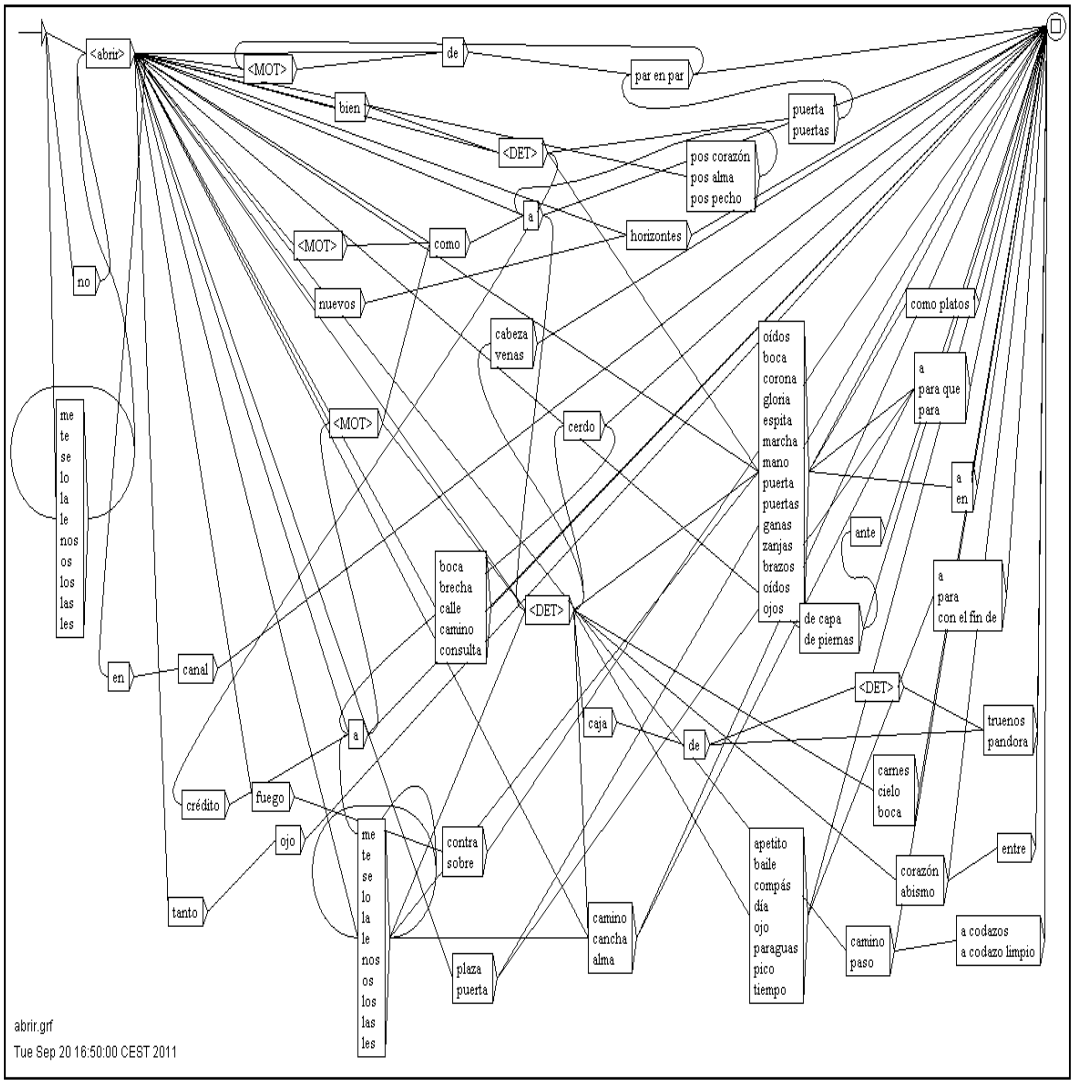

Gráfico nº 17

Unitex detecta la presencia de las realizaciones recogidas en los distintos gráficos reunidos en el transductor dentro del corpus pretratado, que serán destacadas dentro del texto. La búsqueda de los gráficos con el programa en los textos publicados el día 20 de septiembre de 2011, extraídos automáticamente de periódicos (*El Mundo* y *El País*), nos muestra los siguientes resultados.

Los resultados se pueden resumir en el siguiente cuadro:

Verbo	Expresión	Frecuencia
Abandonar	Abandonar el barco	2
	Abandonar el campo	2
	Abandonarse a	2
Abrir	Abrir las puertas (a)	14
	Abrir la puerta (a)	18
	Abrir la boca	2
	Abrir paso	10
	Abrirse paso	2
	Abrir fuego contra	2
	Abrir los ojos	2
	Abrir boca	4
	Abrir brecha	4
	Abrir la caja de los truenos	2
	Abrir ojos	2
	Abrir una/la caja de pandora	4

Cuadro nº 5.

CONCLUSIÓN

Las unidades fraseológicas son una muestra muy representativa de la idiosincrasia de las lenguas y de las culturas que han sido tratadas muy superficialmente por los diccionarios. La informática y las nuevas tecnologías pueden facilitar, a partir de grandes bases de datos, exhaustivos su tratamiento y su estudio por los usuarios nativos y los estudiantes de lenguas extranjeras. Las aplicaciones contempladas por los diccionarios electrónicos son muy numerosas y van a depender en gran parte de la información y del tratamiento que se les haya dado. Pero lo más interesante es que pueden suponer una verdadera revolución en el tratamiento, la selección, la enseñanza y la traducción de estas formas. En efecto, el estudio que deseamos desarrollar no se limita a ofrecer en una aplicación informática los equivalentes fraseológicos de las CVF seleccionadas sino que va a ofrecer varias aplicaciones útiles para los usuarios españoles, para los estudiantes del español lengua extranjera y para los traductores como son las CVF más usadas en la actualidad en las cinco variantes diatópicas que permitirán establecer el núcleo mínimo

competencial fraseológico, sus equivalentes en alemán, árabe, catalán, francés, inglés e italiano. Esta selección de expresiones más usuales será un elemento de gran importancia para la transmisión de la lengua y de la cultura española. Por ello varias Universidades extranjeras han aceptado utilizarla en su docencia.

Estas CVF más empleadas constituirán el mínimo fraseológico que servirá de referente para su enseñanza en nuestras pruebas de validación con las universidades españolas y extranjeras que aceptan de participar en la verificación y la enseñanza del mínimo fraseológico más usado y referenciado sacado a partir de la base contextual.

REFERENCIAS BIBLIOGRÁFICAS

CORPAS PASTOR, G. (1996): *Manual de fraseología española*. Madrid: Gredos.

GARCÍA-PAGE, M. (2008): *Introducción a la fraseología española*. Barcelona: *Estudio de locuciones*. Anthropos.

GONZÁLEZ REY, M. I. (2002): *La phraséologie du français*. Toulouse : Presses Universitaires du Mirail.

GROSS, G. (1996): *Les expressions figées en français : noms composés et autres locutions.* Gap-Paris: Ophrys.

GROSS, M. (1982): «Une classification des phrases figées du français», *Revue Québécoise de Linguistique*, 11.2 : 151-185.

MEJRI, S. (1997): *Le figement lexical. Descriptions linguistiques et structuration sémantique*. Publication de la Faculté des Lettres de la Manouba.

MOGORRÓN HUERTA, P. (2002): La *expresividad en las locuciones verbales en francés y en español*. Alicante: Publicaciones Universidad de Alicante.

MOGORRÓN HUERTA, P. (2004): Los diccionarios electrónicos fraseológicos, perspectivas para la lengua y la traducción. *E.L.U.A.*, nº 12. Cifuentes, JL, & Azorín, D. eds, Universidad de Alicante.

MOGORRÓN HUERTA, P. (2008): «Traduction et compréhension des locutions verbales», *Meta*, 53, nº 2, 378-406.

MOGORRÓN HUERTA, P. (2010): «Analyse du figement et de ses possibles variations dans les constructions verbales espagnoles», *Linguisticae Investigationes*, 33:1 Amsterdam/ Philadelphia: John Benjamins.

RUIZ GURILLO, L. (1997): *Aspectos de fraseología teórica y aplicada*. Universidad de Valencia, anejo 24 de *CF*.

SEVILLA MUÑOZ, J. (1993): « Las paremias españolas: clasificación, definición y correspondencia francesa », *Paremia*, 2: 15-20.

ZULUAGA OSPINA, A. (1980): *Introducción al estudio de las expresiones fijas*. Frankfurt: Verlag Peter Lang, Studia Romancia, nº 10.

Diccionarios

AoMex = *Diccionario breve de mexicanismos*. http://www.academia.org.mx/diccionarios/DICAZ/inicio.htm.
DDDYEDE = *Diccionario de dichos y expresiones del español*. Madrid. Abada. 2011.
DDFE = *Diccionario de fraseología española. Locuciones idiotismos modismos y frases hechas usuales en español* [su interpretación]. Madrid. Abada. 2007.
DEA = *Diccionario del Español Actual*. Madrid: Aguilar lexicografía, 1999.
DHDA = *Diccionario del habla de los argentinos*, Academia Argentina de Letras, Buenos Aires: Espasa Calpe. 2003.
AR= REY, A. & CHANTREAU, S. (1979) : *Dictionnaire des expressions et locutions figurées*. Paris : Larousse.
DFDEA = *Diccionario fraseológico documentado del español actual*. Madrid: Aguilar lexicografía, 2004.
DFDEM = *Diccionario fraseológico del español moderno*. Madrid: Gredos, 1994.
DRAE. *Diccionario de la Real Academia Española*: (vigésimo primera edición). Madrid: Espasa-Calpe, 1992.
DT= *Diccionario temático de locuciones francesas con su correspondencia española*. Madrid: Gredos, 2004.
DTDFH = *Diccionario temático de frases hechas* (2004), de S. Rodríguez-Vida. Barcelona: Columbus.
DUE = MOLINER, M. (1966-67 = 1999): *Diccionario de uso del español.* Madrid: Gredos.
EPM = *Enciclopedia Planeta Multimedia*, edición 2005 en DVD-ROM.
Espasa = Diccion@ario Espasa 2.0 Diccionario de la lengua española. Diccionario de sinónimos y antónimos.
GDEBI =*Grand Diccionario Espasa español-francés / francés-español*. Madrid: Espasa-Calpe, 2000.
GDFHL = *Gran Diccionario de Frases Hechas*. 2001. Barcelona: Larousse.
GDLE = *Gran Diccionario de la Lengua Española*. Barcelona: Larousse, 1999.
LARBI = *Larousse moderno français-espagnol español-francés*. Paris: Larousse, 1993 (1ª ed. 1967).
LBI = *Gran diccionario Larousse Español-Francés / Francés-español* (1999). Barcelona: Larousse.
LIBSA = *Jergas, Argot y Modismos*. Lengua Española. Madrid. 2001.