COMBINATORIA TERMINOLÓGICA Y DICCIONARIOS ESPECIALIZADOS PARA TRADUCTORES [1]

CHELO VARGAS SIERRA [2]
Universidad de Alicante

1. INTRODUCCIÓN

Un problema de traducción que presentan los textos especializados es la terminología. De hecho, se estima que en un porcentaje superior al 40% del tiempo invertido en una traducción técnica o científica se dedica a solucionar los problemas terminológicos inherentes a la traducción de cada texto (Arntz, 1993, Walker, 1993). Dichos problemas pueden ser de distinta naturaleza; por ejemplo, de neología, de equivalencia conceptual o lingüística, de variación denominativa (concurrencia sinonímica), entre otros. Asimismo, durante el proceso de codificación o reformulación del texto meta (TM), el traductor ha de poder solucionar otro escollo: la colocación adecuada en contexto de las unidades terminológicas (UT). En efecto, por lo general el traductor es capaz de encontrar soluciones satisfactorias a problemas terminológicos, pero en el momento de incorporar los términos en su entorno lingüístico y producir un contexto natural en la lengua de especialidad, es decir, tal y como lo redactaría un experto en la materia, es cuando encuentra sus mayores dificultades. Si es clara esta ne-

[1] La investigación realizada para el presente estudio es parte de un proyecto de investigación sobre combinatoria terminológica (ref. n. PR2009-0459), financiado por el Ministerio de Ciencia e Innovación, mediante el Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de I+D+i 2008-2011.

[2] Miembro de los grupos de investigación IULATERM (Lèxic i Tecnologia) del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra de Barcelona (Ref. 2009SGR1306) y 'El Inglés Profesional y Académico' (Universidad de Alicante).

cesidad que experimentan los traductores de poner los términos en contexto la utilidad de los diccionarios combinatorios terminológicos de uso multilingües (DICTUM) está fuera de toda duda, pues proporcionarían la información que permitiría que los traductores combinasen adecuadamente los términos con otras unidades lingüísticas, con las que los primeros mantienen alguna asociación relevante de orden semántico, sintáctico o léxico.

Ilustremos estas asociaciones con unos cuantos ejemplos contrastivos. Pongamos el caso de un traductor especializado en el ámbito de la economía cuyas lenguas de trabajo sean el inglés y el español. Este profesional necesita tener un recurso terminográfico que le permita saber que el equivalente al inglés de 'cheque en blanco' es *blank cheque* (no **white cheque*), un 'cheque cruzado' es *crossed cheque*, pero *open cheque* no se traduce por *cheque abierto, sino por 'cheque (pagadero) al portador', 'cheque sin cruzar' o 'cheque no cruzado'; del mismo modo, en inglés los cheques pueden estar *stale* y no *expired* , es decir, vencidos o caducados, pero no expirados; y que *to write a cheque* se traduce por 'extender un cheque' y no *escribir un cheque.

Para concluir esta introducción, expondremos a continuación las razones que nos han llevado a elegir el término 'combinación' en lugar de otros como 'colocación', 'frasema' o unidad fraseológica.

Al revisar bibliografía relevante sobre combinatoria en los lenguajes de especialidad (Cabré, Lorente y Estopà, 1996; Gambier, 1992; Gläser, 1994; Heid y Freibott, 1991; L'Homme, 2000; Lorente, Bevilacqua y Estopà, 1998; Martin, 1992; Meyer y Mackintosh, 1996, entre otros), hemos constatado que no existe ni un consenso denominativo ni una distinción tajante sobre si una determinada estructura es un término compuesto (TC), una colocación o una unidad fraseológica especializada (UFE). En efecto, hemos encontrado que los mismos ejemplos (*fuentes renovables*, *jet engine* o *house arrest*) son denominados por unos lingüistas 'colocación' y por otros 'unidad terminológica polilexemática' o TC. Lo que Lorente (2002) incluye bajo la denominación de 'unidad fraseológica especializada' (*denunciar un contrato*, *la denuncia del contrato*) otros lo denominan 'colocación' a secas (Koike, 2001), o 'colocación terminológica' en artículos donde se tratan específicamente los lenguajes profesionales y académicos (González Rey, 2002; Ruíz Gurillo, 2002).

Por ello, y adaptando el término a partir del diccionario *Práctico* de Bosque (2006), cuya macro y microestructura nos sirve, además, de modelo, hemos adoptado una denominación única, esto es, la de 'combinación terminológica', que utilizamos como expresión hiperonímica que abarca y alude a todos los tipos de combinaciones léxicas que incluyen al menos dos lexemas —una UT más otro constituyente de distinta categoría gramatical— entre los que operan restricciones de orden semántico, sintáctico o léxico.

La meta de este artículo es doble: por un lado, presentar un estado de la cuestión sobre los trabajos previos que se han llevado a cabo para construir diccionarios multilingües de combinatoria especializada, así como de las herramientas disponibles para automatizar las tareas involucradas en la elaboración de tal recurso; por el otro, mostrar nuestra propuesta metodológica y de entrada terminográfica. Para ello, en el segundo apartado se presentan una serie de proyectos llevados a cabo sobre combinatoria terminológica multilingüe ordenados cronológicamente. A continuación, en el tercer apartado, nos centraremos en una breve presentación y revisión de algunas herramientas de extracción de colocaciones y de concordancias. Tras este estado de la cuestión relativo a proyectos emprendidos y herramientas útiles, en el cuarto se presenta una propuesta metodológica para la elaboración de un DICTUM. El apartado quinto muestra un esbozo de entrada combinatoria especializada bilingüe y su diseño en una base de datos relacional. Cierra este artículo el apartado final de conclusiones.

2. Proyectos terminológicos multilingües de combinatoria

La puesta al día y la síntesis de los trabajos que se publican en un área del saber determinada resultan ser parte fundamental de todo investigador que se precie, pues ello hace posible que la ciencia avance y que la investigación original establezca sus fundamentos en los trabajos previos.

Uno de los objetivos metodológicos que nos marcamos al concebir esta investigación era, precisamente, realizar una actualización de los trabajos previos emprendidos para acometer proyectos multilingües sobre combinaciones terminológicas. Con este objetivo, los pasos realizados han sido, en primer lugar, hacer una extensa búsqueda bibliográfica (accediendo a bases de datos documentales y otra información a través de Internet) y, en segundo, remitir consultas a las listas de distribución *Corpora-list* y *Euralex* con el fin de obtener respuestas de expertos que hubieran trabajado con anterioridad nuestro objeto de estudio.

Es necesario hacer constar que únicamente hemos indagado sobre los proyectos bilingües o multilingües que abordaran el tema de las colocaciones en ámbitos de especialidad, y entre sus lenguas de trabajo tenían que incluir el inglés y/o el español. En consecuencia, no recogeremos aquí los proyectos monolingües que se hayan podido llevar a cabo hasta el momento sobre el lenguaje general, pues supera los límites que nos marcamos para esta investigación. Sin embargo, sí incluimos algunos proyectos bilingües sobre colocaciones (no expresiones idiomáticas ni modismos) del lenguaje general. Esta inclusión se debe a que en las descripciones de estos proyectos hemos encontrado alusiones a la

importancia y necesidad de recoger en sus respectivos recursos combinaciones especializadas.

Otra cuestión que conviene destacar son los diferentes puntos de vista desde los que se abordan las colocaciones. Como es sabido, este fenómeno lingüístico es objeto de numerosos estudios en lexicografía, en lingüística de corpus y en procesamiento del lenguaje natural (PLN). En este último enfoque, encontramos a los expertos en informática, cuyo núcleo investigador gira en torno a múltiples aspectos sobre sus aplicaciones informáticas. Este último enfoque no forma parte intrínseca de esta investigación, pues el marco en el que nos situamos es de orientación más lingüística, si bien recogeremos un catálogo de algunas herramientas desarrolladas de las que hemos tenido conocimiento que nos parecen válidas para acometer alguna de las etapas de procesamiento involucrados en trabajos de combinatoria.

Los proyectos de los que hemos tenido conocimiento son los que referenciamos a continuación y que hemos ordenado cronológicamente.

El primero, cuyo título es *Internet. Répertoire bilingue de combinaisons léxicales spécialisées (français-anglais)* fue desarrollado por Meynard (Meynard, 2000). Su objetivo era la elaboración de un repertorio especializado sobre la combinatoria más usual de términos utilizados en el ámbito de Internet, y las lenguas de trabajo fueron el francés e inglés. Este repertorio se dirige a traductores, redactores técnicos y terminólogos, por la necesidad que tienen de promover el desarrollo y facilitar la comprensión de Internet a pesar de no ser necesariamente especialistas en esta materia. Este repertorio bilingüe presenta una descripción no exhaustiva de los usos del inglés y francés en este campo. El diccionario se articula alrededor de palabras claves (términos base). Se encuentran recogidos los equivalentes de las combinaciones de estos términos con verbos, sustantivos y adjetivos que co-ocurren con aquél. También incluye una definición que ilustra la acepción en la que se emplea cada término base.

El segundo se denomina *English CrossLexica. Diccionario de combinaciones de palabras*, y sus investigadores fueron Bolshakow y Gelbukh (Bolshakow y Gelbukh, 2001). Su objetivo primordial es la elaboración de una base de datos de colocaciones y creación de un sistema inteligente capaz de deducir millones de combinaciones probables no codificadas directamente en el diccionario. Las lenguas de trabajo de este recurso fue el inglés, aunque el repertorio contenía equivalencias al español; inicialmente, se desarrolló la base de datos en ruso con equivalencias al inglés. Se trata de un tesauro y diccionario de combinaciones de palabras en inglés, dirigido a estudiantes y profesores de inglés, así como a informáticos especializados en el procesamiento del lenguaje natural. Además de la combinatoria de sustantivos y verbos, también ofrece los modelos de subcategorización de verbos, sustantivos y adjetivos, las formas morfológicas de pala-

bras, equivalencias de las combinaciones, sinónimos, antónimos, hipónimos, hiperónimos y marcas de uso.

El tercero, *Linguistic Analysis and Collocation Extraction* (Seretan, 2008; Seretan *et al.*, 2004), es un proyecto financiado por la red Geneva International Academic Network (GIAN). Las instituciones que lo llevan a cabo son Laboratoire d'analyse et de technologie du langage (LATL) y Organización Mundial del Comercio (OMC), y su cordinador es el Prof. Eric Wehrli (LATL). Como lenguas de trabajo principales están el inglés y el francés. Posteriormente ha sido ampliado al español y al italiano. El objetivo del proyecto es diseñar y desarrollar un sistema informático de extracción terminológica basado en el análisis lingüístico capaz de gestionar expresiones polilexemáticas. El trabajo se realiza con corpus. Concretamente, en su fase inicial, el análisis y procesamiento se realiza sobre textos paralelos (inglés-francés) procedentes de la OMC. El corpus es procesado por el analizador sintáctico *Fips Syntactic Parser*. Para la extracción de colocaciones el equipo investigador se fundamenta en un sistema híbrido, en el que se emplea información lingüística y estadística. En la primera etapa, los resultados del vaciado los proporciona el componente denominado *Co-occurrences Extraction System*, que identifica todas las co-ocurrencias de dos palabras o bigramas que siguen los patrones adjetivo-nombre, nombre adjetivo, nombre-nombre, sujeto-verbo, verbo-objeto, nombre-preposición-nombre, verbo-preposición, verbo-[preposición]-argumento, definidos previamente. A este primer proceso de vaciado le sigue un test estadístico, que consigue ordenar las colocaciones obtenidas según su importancia colocacional, es decir, en primer lugar situará las combinaciones léxicas con muchas probabilidades de constituirse en una colocación y en los puestos más bajos a aquéllas que tienen una posibilidad menor. El programa presenta al usuario este ordenamiento de los bigramas extraídos junto con los contextos de uso de la colocación en cuestión e información sobre el nombre del fichero. Asimismo, contiene un módulo diseñado para extraer colocaciones de más de dos palabras (*Multi-Word Collocation Extraction Module*). El paquete informático que se desarrolla para extraer las colocaciones del corpus de trabajo incluye módulos de gestión de colocaciones bilingües diseñados para terminólogos y traductores. Estos módulos permiten: (1) visualizar segmentos traducidos; (2) crear y mantener una base de datos de combinaciones terminológicas; (3) identificar una colocación dada en un texto; y (4) ver la traducción propuesta para el segmento meta en donde aparece contextualizada la colocación. Los campos que incluye la entrada son: la colocación, su base y su colocado, el índice de colocaciones del lexicón, el patrón sintáctico, el/los equivalente(s), contexto de uso en la LO y sus contextos paralelos al resto de lenguas.

El siguiente proyecto es el *Diccionario de unidades fraseológicas inglés-español / español-inglés* (Molina, 2006), financiado por el Ministerio de Ciencia y Tecnología (proyecto de I+D BFF02540; duración 2001-2003), y dirigido por Molina. Se desarrolla en el seno del Departamento de Filología Moderna, Facultad de Letras, de la Universidad de Castilla-La Mancha, y sus lenguas de trabajo son el inglés y el español. Su objetivo concreto es compilar colocaciones, expresiones idiomáticas y unidades fraseológicas en inglés por orden semasiológico, junto con sus equivalentes en español. El trabajo sobre corpus se desarrolla con el *British National Corpus* (BNC) y el *Bank of English* para extraer las 10.000 colocaciones más frecuentes. Una vez compilada la colección de colocaciones en inglés proceden con su traducción al español, para la que incluyen 25.000 ejemplos de uso real tomados del *Corpus de Referencia del Español Actual* (CREA).

Un proyecto todavía vigente es el denominado *Collocations en contexte: extraction et analyse contrastive* (Todirascu *et al.*, 2008), que cuenta con financiación de la Agence universitaire de la francophonie (AUF). Son varias las universidades, instituciones e investigadores involucrados (Todirașcu y Gledhill [LILPA, Université Marc Bloch, Estrasburgo, Francia]; Heid y Weller [IMS Stuttgart, Universität Stuttgart, Alemania]; Stefanescu y Tufis [ICIA, Academia Româna, Bucarest, Rumania]; y Rousselot [INSA, Estrasburgo, Francia]). Se marca dos objetivos específicos: (1) el desarrollo de un sistema de extracción semiautomático de colocaciones, capaz de explotar corpus alineados; (2) la constitución de un diccionario multilingüe de colocaciones a partir de los candidatos colocaciones propuestos por el sistema de extracción. Las lenguas de trabajo son el rumano, el francés y el alemán. El inglés se utiliza como lengua intermediaria para la extracción a partir de los córpora alineados. El trabajo se realiza sobre el corpus alineado *AcquisCommunautaire-ACQ*, que consta de 21 lenguas europeas, y es especializado (lenguaje jurídico-administrativo, compuesto de la legislación europea publicada desde 1950). Este corpus se emplea tanto para establecer una lista de colocaciones comunes en los tres idiomas, como para identificar las colocaciones específicas de cada lengua. Como corpus de validación emplean córpora monolingües compuestos por periódicos en francés (*Le Monde* y *Le Monde diplomatique*, 44 m. de palabras aprox.), en alemán (*Stuttgarter Zeitung* y *Frankfurter Rundschau*, 76 m. de palabras aprox.), en rumano varios periódicos y textos médicos (10 m.), y en inglés cuentan con el *British National Corpus Baby* (4 m.) y con el BNC completo. Para etiquetar y lematizar los córpora en inglés, francés y alemán se empleó el *TreeTagger* y para el rumano fue el *Tokenizing Tagging Lemmatizing Free Running Text*. El equipo ha desarrollado su propio extractor de colocaciones (focalizado en las construcciones verbo-nominales). Se trata de un sistema híbrido, esto es, utiliza en primer lu-

gar un sistema estadístico (independiente de las lenguas), seguido de una etapa de filtrado lingüístico (basado en patrones).

Otro proyecto de desarrollo en el ámbito nacional es el *Scie-Lex. Diccionario Electrónico de Combinaciones Léxicas en el Inglés Científico* (Verdaguer *et al.*, 2008), que cuenta con financiación del Ministerio de Ciencia y Tecnología y FEDER. Su coordinación corre a cargo de Verdaguer (Universidad de Barcelona). La base de datos que pretenden construir es monolíngüe (en inglés) con equivalencias del término base al español. Su objetivo: construir una base de datos léxica que proporcione información sobre el uso correcto de patrones sintácticos y colocacionales de palabras no técnicas en el registro científico y sobre las características fraseológicas convencionalizadas de géneros. El trabajo se realiza sobre un corpus etiquetado morfológicamente del ámbito científico, en inglés, y con más de tres millones de palabras, y la extracción se lleva a cabo por medios manuales y semiautomáticos, haciendo uso de los programas *WordSmith Tools* y *Concgram*. La base de datos de combinaciones terminológicas es de elaboración propia, creada en *Access*.

El proyecto *KWiC Web Guide to Medical English for German-Speaking Health Professionals / Fachwortschatz Medizin Englisch. Sprachtrainer und Fachwörterbuch in einem* (Friedbichler y Friedbichler, 2009) surge en el seno de la Innsbruck Medical University, Austria. Su objetivo es la elaboración de materiales didácticos bilingües (inglés-alemán) con información colocacional para términos del ámbito de la medicina y la odontología, y las lenguas de trabajo el alemán e inglés, con un próxima ampliación al danés. Se trata de un diccionario de aprendizaje de lenguas de especialidad basado en corpus de medicina, distribuido en papel y en CD-ROM. Contiene más de 100.000 entradas, que incluyen la combinatoria del término, definición, contextos, marcas de uso (inglés británico o americano), información fonética, abreviaturas, información gramatical, sinónimos, antónimos, términos relacionados, notas, e información equivalente al alemán. El diccionario se divide en ámbitos temáticos. Las palabras clave en inglés, los contextos ilustrativos y las colocaciones se obtuvieron de un corpus de más de 20 millones de palabras compuesto por textos médicos, todos ellos procedentes de fuentes auténticas y profesionales y escritos por autores nativos. El software empleado para realizar la extracción fue *TransConc*, programa de concordancias especialmente diseñado para la edición y la traducción especializada.

El último proyecto que relacionaremos es el *DicoInfo. Le dictionaire fondamental de l'informatique et de l'Internet* http://olst.ling.umontreal.ca/dicoinfo/ (L'Homme, 2009). Cuenta con financiación del Fonds québécois de la recherche sur la société et la culture (FQRSC), y del Conseil de recherches en sciences humaines (CRSH) de Canadá. Su coordinadora es L'Homme y se desarrolla por el grupo de investigación ÉCLECTIK y el Observatoire de Linguistique Sens-Texte (OLST). Las

lenguas sobre las que trabajan son el inglés y el francés, y el objetivo es construir una base de datos léxica que proporcione información léxico-semántica sobre los términos. La base de datos recoge la estructura actancial de los términos —que pueden ser tanto sustantivos, como verbos y adjetivos—, distinciones semánticas muy detalladas y listas de relaciones léxicas compartidas por el término en cuestión con otras unidades léxicas. La entrada incluye hasta 10 categorías de información: categoría gramatical, estatus, definición, sinónimos, información administrativa (fecha de la última actualización y nombre del terminólogo/a), estructura actancial, contextos, relaciones léxicas y realizaciones lingüísticas de los actantes. Las colocaciones aparecen en el campo 'relaciones léxicas' y se listan a partir de un término base, que constituye la entrada de la ficha terminográfica. El corpus de trabajo contiene aproximadamente dos millones de palabras y se compone de textos sobre informática e Internet. La información colocacional que se obtiene de este corpus se comprueba a través de consultas adicionales que se realizan en la web. *TermoStat* es un extractor automático de términos cuya identificación de candidatos se basa en el contraste de corpus especializados con otros de carácter general. La versión disponible en línea de este extractor es la siguiente: <http://olst.ling.umontreal.ca/~drouinp/termostat_web/>. Puede trabajar con textos en los idiomas francés, inglés, español e italiano (beta).

Los trabajos realizados hasta el momento que recogen los resultados de las investigaciones llevadas a cabo sobre la combinatoria terminológica multilingüe demuestran que es bastante reciente el interés que despierta el fenómeno entre los creadores de aplicaciones terminológicas como la que describimos en este trabajo. Constatamos que todos ellos realizan sus análisis y estudios a partir de un corpus constituido, algunos etiquetados. Como es lógico, en los proyectos en cuyos equipos trabajan investigadores de PLN (Todirașcu *et al.*, 2008; Seretan *et al.*, 2004), las herramientas de procesamiento de corpus y de extracción de colocaciones son más avanzadas y específicas para tareas concretas del proceso. El resto hace uso de programas de concordancias comercializados para llevar a cabo la extracción. También hemos verificado que no existe ningún sistema gestor de bases de datos (SGBD) de colocaciones multilingüe como tal disponible en el mercado. De hecho, podemos afirmar que todos los proyectos arriba referenciados han hecho uso de bases de datos creadas *ad hoc*.

3. Aplicaciones informáticas

La estación de trabajo de un terminólogo está compuesta por un buen número de aplicaciones informáticas. Cada una de ellas permite realizar una o varias tareas concretas dentro de las muchas que componen el flujo de trabajo

terminológico (*cf.* Vargas, 2008). Van desde las más sencillas (procesador de textos, hoja de cálculo, Internet, OCR, etc.) hasta las más complejas: etiquetadores de corpus, extractores y gestores de bases de datos terminológicas. Nos centramos en estas últimas, por ser las más específicas.

El catálogo que expondremos para cada tipo de herramienta no pretende ser exhaustivo; entre otras razones, por los límites de extensión establecidos para este artículo. Nuestra intención es simplemente dar orientaciones del tipo de herramientas disponibles y su función para tres de las tareas más importantes en el trabajo terminológico: el procesamiento del corpus, su análisis y el almacenamiento de los datos extraídos en la fase anterior.

3.1. Procesamiento del corpus (etiquetadores morfosintácticos)

Para procesar un corpus hay que transformar de forma previa el texto original de modo que posteriormente podamos acceder a él y extraer el máximo de información posible. Por ello, el corpus de estudio suele pasar por varios procesos, dependiendo del grado de detalle al que deseemos llegar con el etiquetado, proceso que consiste en la asignación automática de etiquetas lingüísticas. El corpus etiquetado o anotado es el compuesto por textos que contienen etiquetas analíticas que explicitan alguno de sus aspectos lingüísticos. Las etiquetas pueden ser morfológicas, que vinculan cada palabra con su categoría gramatical (sustantivo, verbo, adjetivo, etc.), sintácticas (sintagma nominal, verbal, etc.) e incluso semánticas, pragmáticas o discursivas.

El etiquetado más habitual y canónico es el morfosintáctico (*part-of-speech tagging*). Se ha convertido en la forma canónica de etiquetado de un corpus por dos razones: a) porque es lo suficientemente sencillo como para que el etiquetado de la mayor parte del texto se realice de forma automática; y b) porque sus utilidades son obvias, por ejemplo, en el campo de la lexicografía, donde constituye el primer paso para la lematización automática.

El etiquetado del corpus a este nivel suele comprender las fases de: (1) segmentación (en el PLN este proceso recibe el nombre de 'tokenización', que consiste en la segmentación del texto en cadenas de caracteres y cifras que se encuentran entre espacios); (2) lematización (especificación de la forma no marcada de cada palabra) (3) análisis morfológico y categorial (asignación de posibles categorías gramaticales y morfológicas); y (4) etiquetado morfosintáctico o desambiguación de las categorías gramaticales dudosas.

Seguidamente exponemos cuatro etiquetadores morfosintácticos. La selección la hemos realizado en función de su popularidad y disponibilidad actual, bien de forma gratuita o bien mediante pago. Otro parámetro que hemos tenido

en cuenta para dicha selección es que los programas debían funcionar en distintos sistemas operativos, entre los que se encontrara Windows, por ser éste el más extendido hoy por hoy.

CLAWS utiliza también técnicas estadísticas para marcar el corpus. Su última versión (*CLAWS4*) es un sistema híbrido (estadístico y lingüístico) que se empleó para etiquetar el *British National Corpus*. Su página web es: http://ucrel.lancs.ac.uk/claws/. Funciona con los sistemas operativos Unix y Windows. El programa no es gratuito, pero hay una versión demo disponible en línea: http://ucrel.lancs.ac.uk/claws/trial.html.

Freeling es un paquete de programas de análisis lingüístico gratuito, desarrollado por el grupo de investigación TALP, de la Universidad Politécnica de Cataluña. Además de etiquetar morfosintácticamente un corpus, ofrece otras posibilidades de análisis (segmentación por frases, por cadenas de caracteres entre espacios en blanco, reconocimiento de palabras compuestas, entre otras). Su página web es: http://www.lsi.upc.edu/~nlp/freeling/. También es posible procesar texto desde la versión demo en línea (http://garraf.epsevg.upc.es/freeling/demo.php). Este paquete puede trabajar con textos en catalán, español, gallego, italiano e inglés y los resultados de etiquetado son variados: análisis morfológico, etiquetado morfosintáctico, análisis sintáctico superficial (*shallow parsing*) y análisis de dependencias (*dependency parsing*). La plataforma natural de este paquete, según se expone en su página web, es Linux/Unix, pero también es posible utilizarlo con Windows (<http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/ winfreeling.htm>).

QTAG es un etiquetador estadístico de libre distribución para usos no comerciales desarrollado por Oliver Mason (Universidad de Birmingham). Al utilizar técnicas exclusivamente estadísticas para la desambiguación, es independiente de la lengua y, por lo tanto, puede ser adaptado para que trabaje con otras lenguas distintas del inglés. Su página web es. http://phrasys.net/uob/om/software. Está desarrollado en Java y funciona con los tres sistemas operativos más empleados, esto es Linux, Mac OSX y Windows.

TreeTagger es una herramienta estadística para anotar gramaticalmente un texto y lematizarlo. Fue elaborado por Helmut Schmid del Institute for Computational Linguistics (Universidad de Stuttgart). Puede etiquetar corpus en múltiples idiomas, entre los que se encuentran el inglés y el español. Su página web es: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html. Funciona con distintos sistemas operativos, entre los que se encuentra Windows, y es gratuita para fines de investigación y docencia.

3.2. **Análisis del corpus**

Las herramientas de análisis de corpus que a continuación aparecen están divididas en dos grandes apartados. En el primero, exponemos los ocho sistemas de extracción de términos o combinaciones léxicas de los que hemos tenido conocimiento al consultar bibliografía y otra documentación sobre proyectos terminológicos de distinta índole. Seguidamente, en el apartado 3.2.2, se define el funcionamiento básico de los programas de concordancias y referimos algunas de las herramientas más empleadas de este tipo.

3.2.1. ***Sistemas de extracción / detección de términos o combinaciones léxicas***

Estopà (1999: 39) clasifica los mecanismos de extracción de términos que utilizan los detectores automáticos en tres tipos:

a) mayoritariamente estadísticos: utilizan información estadística y, en consecuencia, independiente de la lengua, esto es, hacen uso de criterios de frecuencia y miden el grado de asociación entre las palabras de un término potencial;
b) mayoritariamente lingüísticos: utilizan información lingüística y, por tanto, dependiente de la lengua, para determinar la probabilidad de que una palabra sea un candidato a término;
c) híbridos: combinan información diversa, como estadística, morfológica, sintáctica, y semántica.

El principal problema al que se enfrenta cualquiera de estos sistemas aludidos es diferenciar el término (unidad léxica o grupo de unidades de contenido especializado) del no término (palabra o grupo de palabras pertenecientes a la lengua general). Los detectores mayoritariamente estadísticos realizan diferentes cálculos probabilísticos para llegar a los resultados que ofrecen. Así, miden la frecuencia de aparición absoluta o relativa de una palabra o grupo de palabras, el grado de asociación entre dos unidades (por medio del test estadístico denominado 'información mutua'), el grado de confianza que permite afirmar que existe una asociación de dos palabras (con el test T-score), la calidad de la palabra clave o *keyness* (mediante el logaritmo de la verosimilitud [de *log likelihood*]), por citar unos pocos. La gran desventaja de estos programas es que producen muchos datos no válidos o ruido, por lo que requieren una mayor dedicación humana después de haber obtenido los listados de palabras.

El principio que subyace tras los sistemas mayoritariamente lingüísticos es que, como su propio nombre indica, utilizan información lingüística para deter-

minar la terminologización (*termhood*) de un posible término. Así, son el resultado de la formalización de una serie de planteamientos teóricos acerca de la naturaleza lingüística del término. Estos programas se nutren, básicamente, de alguno de los tres principios siguientes: (1) patrones terminológicos (p. ej., formantes cultos); (2) análisis sintáctico; y (3) información semántica. Una desventaja que suelen presentan estos sistemas es que se corre el riesgo de perder datos válidos o, dicho de otro modo, generan silencio.

La característica más interesante de los sistemas híbridos es que combinan información de distinta índole para producir sus resultados. Así, utilizan información estadística, morfológica, sintáctica, semántica, y algunos también incluyen ontologías.

Tras esta breve definición de los distintos tipos de extractores que existen, relacionamos en los siguientes párrafos varios extractores de terminología y de combinatoria.

Xtract (Smadja, 1993) es un sistema que funciona en tres fases. En la primera, el programa, mediante el uso de métodos estadísticos, recupera bigramas que no aparecen necesariamente de manera contigua en el texto. En la segunda fase, identifica combinaciones multipalabra y expresiones complejas. Haciendo uso de una combinación de métodos lingüísticos y estadísticos, en la tercera fase *Xtract* etiqueta y filtra las colocaciones recuperadas en la primera fase.

Termight (Dagan y Church, 1994) es una herramienta semiautomática para identificar términos y sus equivalentes en textos paralelos. Para realizar la extracción de candidatos a término (mono y poliléxicos) el programa trabaja en dos fases, a las que los autores denominan «tarea monolingüe» y «tarea bilingüe». En la primera, *Termight* emplea un etiquetador gramatical para elaborar un listado con los candidatos a término; a continuación el terminólogo debe filtrar de manera manual este resultado y construir un listado depurado. Para que se pueda llevar a cabo el filtrado, el programa presenta los contextos de uso a través de líneas de concordancias. En la segunda, el programa identifica los equivalentes basándose en una alineación a nivel léxico; para ello, emplea una lista de términos originales y un corpus bilingüe alineado palabra a palabra con *word-align* (Dagan *et al.*, 1993).

Champollion (Smadja *et al.* 1996) es un extractor de colocaciones paralelas que ha sido realizado a partir de *Xtract*. Tomando como base el corpus *Hansard* (compuesto de las actas oficiales de las sesiones del Parlamento canadiense en inglés y francés canadiense), encuentra las colocaciones de los textos paralelos en francés basándose en métodos estadísticos.

TRUCKS (Maynard, 1999) es un extractor de términos poliléxicos que hace uso de diferente tipo de información contextual, esto es, se nutre de información sintáctica, semántica, terminológica y estadística. Está adaptado para

un subdominio de la medicina, esto es, la patología ocular. Su funcionamiento es secuencial: en primer lugar utiliza información lingüística para realizar un listado inicial de candidatos; a continuación aplica a cada uno de ellos una serie de medidas estadísticas para, seguidamente, tener en cuenta otros términos potenciales que aparecen en el contexto léxico del candidato y, finalmente, observa el contexto semántico del candidato.

YATE (Vivaldi, 2001) es un sistema que extrae candidatos a término a partir de un corpus de textos especializados para los ámbitos de la economía, la medicina y el genoma en catalán y español. Sus características más relevantes es que YATE utiliza métodos híbridos (lingüísticos y estadísticos), además de información semántica y ontológica. Su página web es: http://igraine.upf.edu/cgi-bin/Yate-on-the-Web/yotwMain.pl.

SketchEngine (Kilgarriff *et al.*, 2004) es sistema web de explotación de corpus que muestra concordancias (*Concordance*), listados de palabras (*Word List*), combinatoria, a través de la presentación de un mapa de relaciones gramaticales que mantiene la palabra interrogada con otras palabras (*Word Sketch*), sinónimos (*Thesaurus*) y similitudes/diferencias de combinatoria de dos lemas (*Sketch Diff*). Su página es: http://www.sketchengine.co.uk/.

Ngram Statistics Package (Pedersen & Banerjee, 2003): Se trata de un conjunto de programas de libre distribución en lenguaje Perl que permiten estudiar los n-gramas presentes en uno o más ficheros de texto. Utiliza diferentes pruebas estadísticas estándar de asociación (test exacto de Fisher, logaritmo de verosimilitud, prueba de chi-cuadrado de Pearson, coeficiente de Dice, etc.). Se puede acceder a más información y a la descarga del programa desde el siguiente vínculo: http://www.d.umn.edu/~tpederse/nsp.html.

Mercedes (Araya y Vivaldi, 2004) es un detector de términos con dos versiones: una para ser aplicada en textos del *Corpus Técnico del IULA* y otra, consultable a través de internet (http://brangaene.upf.es/proves/mercedes/index.htm) para aplicarlo sobre frases sueltas.

3.2.2. ***Programas de concordancias***

Son múltiples los programas de análisis textual [3] —más popularmente conocidos por 'programas de concordancias'—, gratuitos y comerciales que están

[3] Por las limitaciones de extensión, no podemos incluir y describir aquí cada una de ellos. Remitimos al lector, no obstante, a la página web de David Lee dedicada a la Lingüística de Corpus (<http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm>)

a nuestra disposición[4]. Las concordancias son instrumentos consolidados ya como indispensables en el estudio de las colocaciones y patrones léxicos; por ello, resulta una pieza clave en toda investigación basada en corpus. El funcionamiento de la mayoría de este tipo de herramientas consiste en segmentar el corpus de estudio y ofrecer los datos resultantes en forma de:

a) lista con estadísticas (palabras que contiene el corpus en su conjunto, por texto, número de palabras diferentes, etc.);
b) listado monoléxico ordenado alfabéticamente y/o por frecuencia;
c) listado poliléxico de todas las palabras del corpus o de una selección del usuario. Estos listados pueden ofrecerse a través de agrupamientos léxicos (clusters) o, si el programa tiene la opción, pidiéndole que calcule la información mutua, o cualquier otra medida estadística que incluya de este tipo (Z Score, MI3, Log Likelihood, etc.);
d) líneas de concordancias o listados de aparición de una palabra específica —llamada 'palabra de búsqueda', 'palabra base' y también 'palabra clave', que puede estar formada por una unidad, varias o parte de ésta— acompañada del texto que la rodea (co-texto). El tipo de concordancia más común es Key Word In Context (KWIC) o palabra clave en contexto. Una lista KWIC agrupa las apariciones de la palabra de búsqueda, que aparece destacada en el centro, lo cual permite analizar y detectar con rapidez sus colocadores o palabras que aparecen en su entorno
e) listado de palabras claves (keywords): esta función contrasta una lista de palabras del corpus de estudio con otra lista procedente de un corpus de referencia. El resultado de esta comparación es una nueva lista de palabras clave o palabras cuyas frecuencias son estadísticamente diferentes en el corpus de estudio con respecto al corpus de referencia.
f) gráfico de distribución de la palabra de búsqueda (plot); esta opción permite apreciar de forma visual la posición donde aparece y se repite una determinada palabra a lo largo de todo el texto. El resultado se asemeja a un código de barras que habrá de ser interpretado por el investigador (relevancia de la posición, mayor o menor frecuencia de aparición en una determinada parte del texto, etc.)
g) lista de colocados (collocates): listado de palabras que aparecen alrededor de la palabra base, en posiciones determinadas.

[4] Por las limitaciones de extensión, no podemos incluir y describir aquí cada una de ellos. Remitimos al lector, no obstante, a la página web de David Lee dedicada a la Lingüística de Corpus (<http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm>)

ConcGram 1.0 (Greaves, 2009) es un programa para identificar automáticamente la variación fraseológica. Su funcionamiento básico consiste en encontrar todas las co-ocurrencias de una palabra en el texto (a las que denomina 'concgrams'). El usuario, a través de la observación de las líneas de concordancias, decide si esta combinación constituye una colocación.

WordSmith Tools (Scott, 2008) es, probablemente, en la actualidad el programa de concordancias comercializado más utilizado en la investigación lingüística; prueba de ello son los numerosos artículos, capítulos de libro y tesis doctorales en las que se ha utilizado o se describe de algún modo este programa. Se trata de un paquete informático compuesto de tres herramientas (*Wordlist*, *Concord* y *KeyWords*). Con *WordList* es posible generar listados poliléxicos; de dos palabras, de tres, de cuatro, hasta un total de ocho. Para ello, es necesario realizar primero un índice y a partir de aquí es posible generar bien un listado de la totalidad del corpus de agrupaciones léxicas (*clusters*), o bien pedirle al programa que calcule la información mutua. Con esta última opción el resultado es un listado en donde además de los índices de frecuencia, de la proximidad de las palabras que pone en relación, las veces que aparecen juntas, entre otros datos, muestra una variedad de relaciones colocacionales; concretamente, MI, *Z Score*, MI3 y *Log Likelihood*.

AntConc (Anthony, 2004) es un programa de concordancias gratuito, operativo en sistemas Linux, Mac y Windows. Contiene un conjunto de herramientas para producir líneas de concordancias, listados de palabras y palabras clave por frecuencia, listas de colocados y agrupaciones de una palabra base y gráfico de distribución. Se pueden descargar sus distintas versiones desde: http://www.antlab.sci.waseda.ac.jp/software.html.

Tras haber relacionado y definido brevemente los programas informáticos útiles para el proceso de extracción de combinatoria, a continuación exponemos la propuesta metodológica para la elaboración de DICTUM.

4. Propuesta metodológica para la elaboración de Dictum

Con el fin de lograr los objetivos pretendidos, nuestra propuesta se sustenta sobre tres elementos clave en cuanto a su metodología: (1) sistematizar todo el proceso; (2) trabajar con corpus electrónicos; y (3) utilizar herramientas informáticas que nos ayuden a llevar a cabo el proyecto.

Por 'sistematizar' aludimos a «organizar según un sistema» (RAE). Expresada la anterior definición con más concreción y aplicándola a la actividad práctica de la terminología combinatoria, la sistematización implica articular, en forma ordenada y metódica, todos los elementos —fases, tareas y herramientas—

que componen dicha actividad. Un trabajo terminológico sistemático bilingüe, como el que requiere la elaboración de DICTUM, ha de seguir un proceso constante distribuido en fases (IULA, 2006). Las fases de trabajo que hemos determinado son las siguientes: (1) la definición del trabajo; (2) la preparación; (3) el diseño y construcción del corpus bilingüe del ámbito de estudio; (4) la explotación de corpus para la extracción de combinaciones terminológicas; (5) creación de una base de datos; (6) la aplicación, que incluye el desarrollo integral del diccionario, un proceso de revisión y supervisión del trabajo, más otro final de edición.

El principio metodológico más relevante en el que se fundamenta nuestra propuesta es el de 'hábitat natural' (Cabré, 1999). Partimos de la hipótesis de que un DICTUM debe crearse a partir de un corpus de textos profesionales y académicos específicos del ámbito técnico objeto de estudio. Su utilidad se fundamenta en que lo entendemos como un «ecosistema», siendo las diferentes unidades léxicas y otros elementos las especies que habitan en él. Las palabras pueden estudiarse también *in vitro*, pero si se quiere conocer el modo en que se comportan se deben observar *in vivo*, en su entorno natural, y el entorno natural de las unidades léxicas son los textos. Así, una metodología de trabajo de naturaleza descriptiva y con bases lingüísticas y comunicativas, como la postulada por la Teoría Comunicativa de la Terminología (*ibid.*, 1999), debe valerse, en nuestra opinión, de un método empirista. En un contexto de trabajo terminológico basado en corpus, la adopción de este método significa dar primacía a la observación de los datos lingüísticos en una situación real de uso, reunidos, precisamente, en forma de corpus.

A la hora de realizar la extracción (semi)automática de combinaciones terminológicas, pensamos que la asociación de técnicas estadísticas con información lingüística constituye un criterio fundamental para el reconocimiento de estas unidades. En consecuencia, la situación ideal es utilizar un extractor híbrido. Sin embargo, en ausencia de éste, pensamos que el corpus debe sufrir, al menos, un proceso de etiquetado morfosintáctico; así, cada unidad del corpus contendría información explícita de su categoría gramatical, lo cual permite extraer listados por patrones morfosintácticos con un programa de concordancias como *WordSmith Tools*, además, obviamente, de por frecuencia.

Como método de extracción de listados por patrones morfosintácticos emplearemos, para el idioma inglés, las expuestas en Benson *et al.* (1986); las de Koike (2001) nos servirán como punto de referencia para el español. Sirviéndose de criterios semánticos y sintácticos, los primeros distinguen siete tipos de colocación léxica en inglés, que identifican con los códigos L1, L2, L3... L7, según exponemos con ejemplos escogidos de distintos ámbitos de especialidad:

L1) V (creación y/o activación) + (prep) + N:
- ❒ Creación*: come to an agreement, compose music, reach a verdict*
- ❒ Activación*: set an alarm, fly a kite, launch a missile*

L2) V (erradicación y/o anulación) + N:
- ❒ *reject an appeal, reverse a decision, revoke a license*

L3) ADJ (o nombre utilizado adjetivalmente) + N:
- ❒ *jet engine, house arrest, chronic alcoholic*

L4) N + V (indicando una acción característica de la persona o cosa designada por el nombre):
- ❒ *bombs explode, blood circulates*

L5) Unidad asociada con un N + N (n1 of n2):
- ❒ *a colony of bees, an article of clothing, an act of violence*

L6) ADV + ADJ:
- ❒ *deeply absorbed, strictly accurate*

L7) V + ADV:
- ❒ *affect deeply, anchor firmly*

En lo que respecta al español, Koike (2001: 46) propone seis grupos de colocaciones, que listamos a continuación con nuestros ejemplos:

A) sustantivo + verbo:
- A1) $\text{sustantivo}_{\text{sujeto}}$+verbo:
  - ❒ *circular {la sangre}*
- A2) verbo+$\text{sustantivo}_{\text{cd}}$:
  - ❒ *cometer homicidio, contraer matrimonio*
- A3) verbo+preposición+sustantivo:
  - ❒ *refrentar con mortero, poner (algo) en práctica*

B) sustantivo+adjetivo:
- ❒ *cáncer galopante, capital líquido, daño irreparable, infarto fulminante, precio sucio*

C) sustantivo+de+sustantivo:
- ❒ *banco de peces, juego de herramientas, fajo de billetes*

D) verbo+adverbio:
- ❒ *fallar irrevocablemente, practicar ilegalmente, cerrar herméticamente, cimentar firmemente*

E) adverbio+adjetivo/participio:
- ❒ *diametralmente opuesto, visiblemente afectado, gravemente enfermo, totalmente inmunizado*

F) verbo+adjetivo:
- ❒ *resultar ileso, salir/resultar indemne*

Centrando nuestra atención en las herramientas informáticas, cuatro grandes conceptos de naturaleza tecnológica han protagonizado el panorama terminográfico en esta última década: Internet, los córpora electrónicos, las herramientas para su gestión y extracción terminológica, y las bases de datos terminológicas. La estación de trabajo de un proyecto para realizar DICTUM ha de estar compuesta, en consecuencia, por todas estas aplicaciones tecnológicas necesarias para automatizar las distintas tareas del proceso. El gráfico que sigue pretende ilustrar el flujo de un proyecto DICTUM haciendo uso de instrumentos informáticos específicos para: (a) la gestión y procesamiento de corpus; (b) la extracción de combinatoria especializada; y (c) su almacenamiento y gestión:

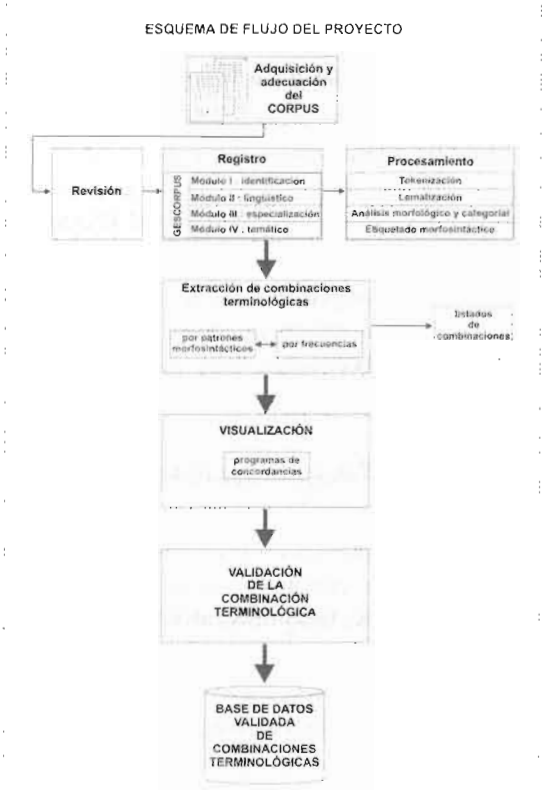

Figura 1: *Flujo de un proyecto de DICTUM*

5. Modelo de entrada terminográfica de combinaciones terminológicas

Crystal (1985: 110) define una 'entrada' como «a term used in semantics to refer to the accumulated structural information concerning a lexical item as

formally located in a lexicon or dictionary», y un diccionario es visto como un conjunto de entradas léxicas. La entrada forma parte de la macroestructura (el eje paradigmático o vertical) y de la microestructura (eje sintagmático u horizontal).

Las entradas de un DICTUM son especiales por la información lingüística que han de contener en el eje sintagmático. Dicha información podemos dividirla en dos planos: (1) las combinaciones más frecuentes de la unidad léxica que encabece la entrada (el lema) con otras palabras; (2) los equivalentes a la otra lengua tanto del lema principal, como de la combinación resultante de combinar dicho lema o base con todos sus colocadores. En el ámbito de los lenguajes de especialidad, la base es generalmente la unidad terminológica y el colocativo es cualquier categoría sintáctica (sustantivo, adjetivo, verbo, adverbio, preposición) que se combine con este término.

Consideraremos a continuación los tipos de categorías de información que un diccionario bilingüe de combinaciones especializadas debería contener de forma ideal. Para ello, adaptaremos la desiderata ya realizada por Haas (1962: 45) a un DICTUM:

1. Debe proporcionar para cada término y su combinación en la LO la equivalencia exacta en la(s) LM incluyendo, asimismo, y de manera especial, el término o combinación que recoge el fragmento que se está traduciendo en el momento de la consulta; por ello puede tener que ofrecer más de un equivalente cuando así lo exija el lema o la combinación:
   a. *deferred bond _ bono de interés diferido / bono de cupón cero / título diferido.*
   b. *falta de lesiones _ minor assault / petty assault / minor batery / petty batery / affray.*
2. Debe contener todos los términos, así como las posibles combinaciones que el traductor necesita consultar junto con sus equivalencias a los idiomas del diccionario;
3. Debe albergar toda la información semántica, conceptual, gramática, temática de dominio y subdominio(s), de flexión y derivación que el traductor pueda necesitar sobre un término dado tanto en la LO como en la(s) LM;
4. Debe recoger información pragmática y de uso, como marcas de frecuencia, de variación diatópica (o geográfica), diacrónicas, diafásicas (o de registro) diastráticas (o socioculturales) y dianormativas, tanto en la LO como en la(s) LM;
5. Debe incluir ejemplos reales de uso procedentes del corpus de vaciado. Dichos contextos han de ser ilustrativos del uso de un término dado y de sus combinaciones;

6. Debe contener toda la información necesaria sobre la ortografía correcta en la LO y en la LM, con observaciones sobre los errores ortográficos más comúnmente realizados, si fuera el caso;
7. Debe proporcionar toda la información fonética que el traductor necesita a fin de que pueda pronunciar correctamente, de un modo semejante a como lo haría un nativo, los términos y sus coocurrentes en las lenguas del diccionario.
8. Debe ser un diccionario equilibrado en cuanto al número de entradas en las direcciones lingüísticas que se aborden (dos, tres, etc.);
9. Debe ser un diccionario que esté orientado tanto para el lector de la LO como de la(s) LM.
10. Debe concebirse como un recurso electrónico, preferentemente, por las posibilidades y rapidez de explotación y de consulta que este formato ofrece frente al tradicional papel.

Tras haber referenciado los anteriores supuestos sobre lo que sería deseable para hacer de un DICTUM una herramienta de trabajo ideal para un traductor, nos queda ahora la difícil tarea de diseñar una entrada terminológica que se adapte a tales exigencias. A continuación presentamos un primer boceto sobre el que hemos estado trabajando. Como se podrá apreciar en el ejemplo, las entradas constan de dos partes. La primera recoge las principales acepciones de la unidad léxica en cuestión y la segunda parte ofrece la combinatoria más frecuente de las distintas acepciones agrupadas por categorías gramaticales.

ACEPCIONES:

authority[1] *n.* gral autoridad, poder, potestad, competencia, jurisdicción, facultades *The solicitors should have invited your client to provide them with the authority to write to the bank.*

authority[2] *n.* proc autoridad, precedente, fuente de prestigio jurídico; doctrina legal o científica; dominio; cita a leyes —*statutes*—, normas —*rules*—, reglamentos —*regulations*—, resolución judicial —*judicial decision*—, libros de texto, etc. *The lawyer grounded his case on recent authorities.*

authority[3] *n.* admin organismo público, organismo autónomo, entidad, ente público, servicio, agencia estatal, junta *He is the president of the airport authority.*

COMBINATORIA:

Combinatoria de la acepción 1:

❑ CON ADJS.

1. absolute / complete / full / supreme / unquestioned ~ (absoluta, completa, plena, suprema, incuestionable, innegable *Attendance by legal representative will suffice for this purpose but that person must be conversant with the case and have *full authority* to act*)

2. governmental / judicial / legal / ministerial / presidential / parental ~ (gubernamental, judicial, legal, ministerial, presidencial, de los padres / paterna *Mrs. Milnes was clear in her own mind that *parental authority* was being usurped*).

3. public ~ (poder público *The ordinary citizen can complain of unlawful action by a *public authority**).

4. control ~ (de control *An airline or even government operated air traffic *control authority* might amend a procedure promulgated in its manuals*).

5. moral ~ (autoridad moral *Although delays were mitigated and judicial efficiency improved, the courts continued to exercise little *moral authority**).

❑ CON VBOS.

1. assert one's / establish / demonstrate / show / exercise / exert / wield / use / have ~ (imponer, instaurar, demostrar, mostrar, ejercer, hacer valer, detentar, usar, tener *Mr Sharif 's immediate aim is to *assert his authority* by getting control of the provincial assembly in Punjabk*).

2. assume ~ (asumir, hacerse con *During the gold-rush of 1858 he (...) assumed authority over the mainland, and sternly maintained law and order in the face of social upheaval*).

3. give (sb) / delegate ~ (dar, conceder, delegar *Authority is delegated to an individual by his manager*).

4. invoke ~ (invocar *State officials invoke authority to enter private property and inspect earthen dams across the state after disaster on Kauai*).

5. give up / relinquish / yield ~ (dejar, ceder, entregar *Teachers relinquish their authority as truth-providers and assume the authority of facilitators*).

6. challenge / defy / deny / rebel against / reject / undermine ~ (desafiar, desobedecer, negar, rebelarse contra, rechazar, minar/socavar *She had *challenge my authority* once too often*).

7. usurp ~ (usurpar *It was not for the courts to *usurp authority* pr operly belonging elsewhere*).

8. abuse / overstep one's ~ (abusar de *It' s very unfair that his boss should be able to *abuse his authority* and make his life so miserable*)
9. exceed one's ~ (*frml*) (excederse en el ejercicio de sus atribuciones, *frml In making these thr eats, Mr Sloan *exceeded his authority**)

❑ CON PREPS.
1. ~ **for** (para *He assumed *authority for* o verseas operations*).
2. ~ **over** (sobre *A commanding officer has complete *authority over* her personnel*).
3. in ~ (los que tienen la autoridad, los que mandan I need to talk to someone in authority).
4. without ~ (sin permiso/autorización She took the car without authority).
5. under the ~ of / under sb's ~ (bajo la autoridad de (alguien) *Many people regard themselves as under the authority of the state*).

❑ EXPRESIONES
1. an air of ~ (un aire de *There was an air of authority about her*).
2. in a position of ~ (tener autoridad *She holds a position of authority in the local church*).
3. power and ~ (poder y autoridad *The central trade union body attained much greater power and authority than either the TUC in Britain*).

Combinatoria de la acepción 2
❑ CON ADJS.
1. appropriate / competent / reliable ~ (apropiada, competente, fidedigna *The appropriate authority was the Attorney General, who retains the power to petition in certain circumstances*).
2. indisputable / irrefutable / unimpeachable / unquestioned ~ (indisputable, irrefutable, indiscutible, incuestionable *The tone of the interviewer was respectful, as if Bala Usman were an unquestioned authority on all matters concerning international relations*).
3. leading / respected / world ~ (importante, principal, de prestigio, respetada, mundial *Sheriff G. N. was a leading authority on sentencing in Scottish courts*).

❑ CON VBOS.
1. to cite / invoke an ~ (citar, invocar, acogerse a *It is unnecessary to *cite authority* to show that both of these orders declaring military zones were illegal and contrary to international law*).

❑ CON PREPS.
1. ~ on (en *Ne wman was a *leading authority* on linguistics*).

❑ EXPRESIONES.
1. the greatest living ~ (la ~ / el/la experto/a más importante vivo *He is regarded as *the greatest living authority* on the affairs of the Middle East*).

Combinatoria de la acepción 3
❑ CON ADJS.
1. district / local / regional ~ (de distrito, local, regional *It is difficult to know how much money is being provided, and whether a *regional authority* is using the money allocated for transport in the block grant for transport matters*).
2. government / state / public ~ (gubernamental, estatal, pública *It is open to a company to complain that a competitor has received a subsidy from a public authority which is incompatible with the Treaty of Rome*).
3. education / health / tax / military / planning / fiscal / port ~ (educativa, sanitaria, (administración) fiscal/administración tributaria/hacienda pública, militar, urbanística, administración/agencia tributaria, portuaria *The defendant dock company, a wholly owned subsidiary of a *port authority*, was granted a long lease*).
4. competent, relevant / lawful / judicial / statutory ~ (competente, legítima, judicial/órgano jurisdiccional/judicial, oficial *The Department of Health and Social Security handed over all its responsibilities for the new tribunals to a newly created *statutory authority* known as the Office of the President of Social Security Appeal Tribunals*).

❑ AUTHORITY + VBOS
1. ~ agree sth/ claim sth / decide sth / deny sth / promise sth (acordar, reivindicar, resolver, negar, prometer *The local *authority denied* negligence*).
2. ~ allow (sb) sth / give (sb) sth / grant (sb) sth (permitir, otorgar, conceder *The local *authority has not granted* planning permission*).

Como indicábamos en el último punto de la desiderata anteriormente expuesta, un diccionario como el que hemos expuesto debería concebirse como un recurso electrónico, por las ventajas obvias que ofrece. La manipulación, almacenamiento y recuperación de los datos terminológicos se lleva a cabo utilizando los sistemas informáticos de gestión terminológica, también denominados 'sistemas gestores de bases de datos terminológicas' (SGBDT), o, de forma

sintética, 'bases de datos terminológicas' (BDT). Son sistemas informatizados de almacenamiento y gestión de unidades terminológicas que se estructuran de acuerdo con determinados criterios, con los usuarios y con la finalidad de la compilación terminológica (*cf.* Vargas, 2008). Los diferentes tipos de SGBDT disponibles en el mercado se diferencian según la complejidad que admita la estructura de las entradas, el número de lenguas con las que pueden trabajar o que se pueden visualizar a la vez, en la flexibilidad para que el usuario pueda definir una determinada estructura de la entrada, entre otros aspectos.

En consonancia con lo apuntado por Vidal (2007: 479), pensamos que una base de datos relacional se configura como un recurso electrónico eficaz y flexible para organizar y registrar un término dado con su información asociada (conceptual, contextual, sinonímica, gramatical, combinatoria, etc.). Este tipo de base de datos, además, permite recuperar la información almacenada en razón de distintos criterios gracias a las relaciones que se establecen entre las múltiples tablas que puede contener.

Las tres figuras que se presentan a continuación proporcionan una visión sintética y provisional de cómo quedaría estructurada la entrada combinatoria en una base de datos y cuáles serían sus relaciones básicas.

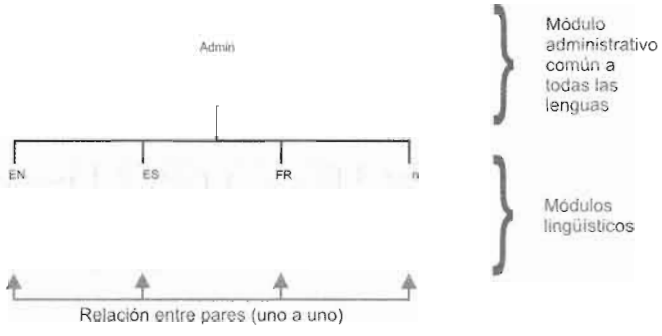

Figura 2: *Estructura de la entrada (módulos administrativo y lingüísticos)*

La estructura de la entrada esbozada en la figura anterior es la habitual en las bases de datos terminológicas disponibles en la actualidad. La entrada tiene dos tipos de datos: los de cabecera, de primer nivel, que es común a todas las lenguas de trabajo y es donde se registran datos de tipo administrativo (nombre del proyecto, número de concepto, fecha de creación...); y los lingüísticos, que contienen toda la información de un término en una determinada lengua.

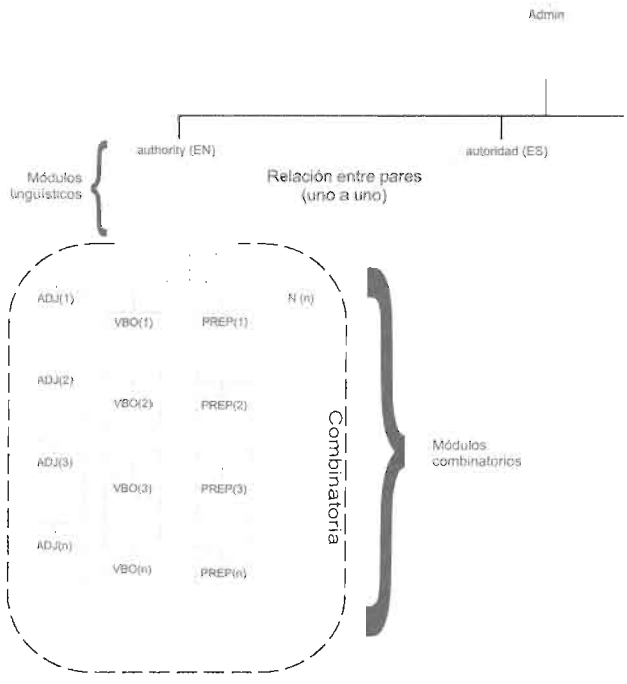

Figura 3: *Estructura de la entrada (módulos lingüísticos y combinatorios)*

Como se aprecia en la Figura 3, los módulos lingüísticos establecen una relación interna entre los lemas de las distintas lenguas de trabajo. Si el módulo administrativo lo entendemos de primer nivel, los lingüísticos quedarían en el segundo. La novedad de este boceto de entrada terminológica combinatoria se encuentra, como es lógico, en los módulos combinatorios, o de tercer nivel. En ellos podrán crearse tantas subfichas como sean necesarias, y se almacenan en cascada teniendo en cuenta la categoría combinatoria a la que pertenezcan (verbo, adjetivo, adverbio, etc.) y siempre dependerán del módulo lingüístico con el que combinan.

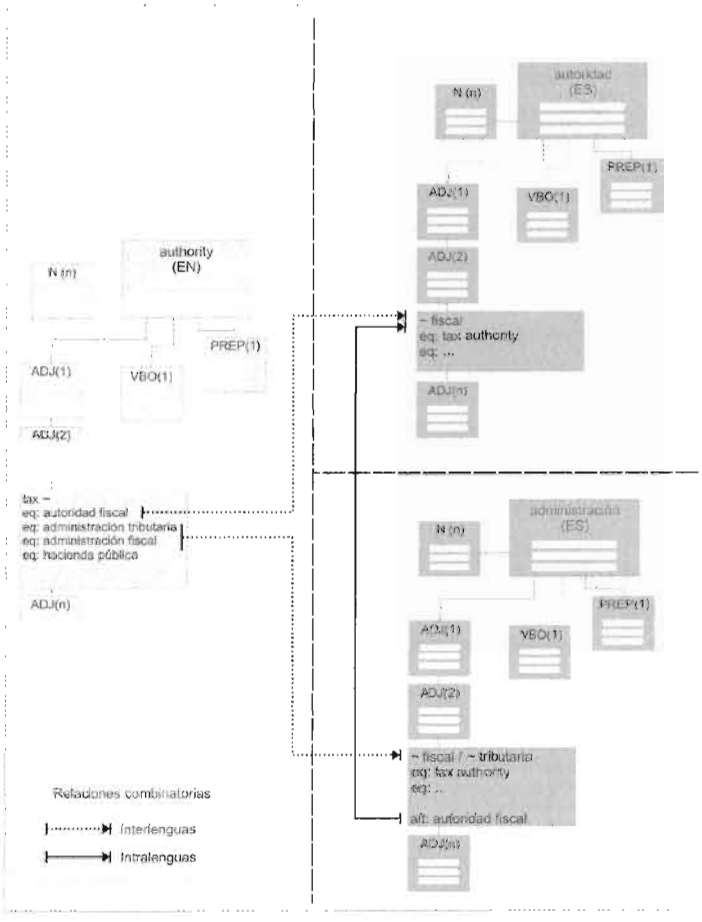

Figura 4: *Relaciones inter e intralenguas de las fichas combinatorias*

Las relaciones entre las fichas pueden producirse tanto en el nivel segundo (módulo lingüístico) como en el tercero (módulo combinatorio). Las relaciones correspondientes al módulo lingüístico pueden ser de tres tipos:

—Relación uno a uno (un lema tiene únicamente un equivalente en otra lengua).
—Relación uno a varios (un lema tiene varias equivalencias en otra lengua).

En la figura 4 hemos intentado plasmar las relaciones inter e intralingüísticas que se dan en los módulos combinatorios. La relación interlenguas corresponde a la relación entre una subficha combinatoria (A) y otra(s) ficha(s) combinatoria(s)

del mismo o distinto módulo lingüístico (B o B-C o B-n). La relación intralenguas es la que se origina entre distintos módulos combinatorios de lemas diferentes en un idioma.

6. Conclusiones

El primer objetivo que nos planteamos en este trabajo era presentar un estado de la cuestión sobre los proyectos terminológicos multilingües emprendidos que abordasen la combinatoria. Las conclusiones que se derivan de la investigación realizada sobre este punto es que las combinatoria terminológica ha despertado un interés investigador relativamente reciente en lo relacionado con la puesta en marcha de proyectos terminológicos encaminados a elaborar recursos que recojan estas piezas lingüísticas, muy valiosas para la traducción. Todos estos proyectos que hemos relacionado basan sus estudios en corpus, algunos de los cuales se encuentran etiquetados. Asimismo, en grupos investigadores en donde están involucrados informáticos expertos en PLN, como era el caso de Todiraşcu et al., 2008; Seretan et al., 2004, L'Homme 2009, se trabaja con herramientas informáticas muy avanzadas, la mayoría de las cuales son de desarrollo propio, pero que no se encuentran disponibles fuera del grupo que los crea. Asimismo, hemos constatado que no existe en el mercado ningún SGBDT específico para combinatoria multilingüe, con lo cual han de crearse *ad hoc* para cada proyecto concreto. Seguidamente, hemos mostrado una relación comentada de aplicaciones informáticas disponibles para emprender este tipo de trabajos, dividida dicha relación en dos grupos: herramientas para procesar el corpus, y las empleadas para analizarlo. Los resultados de esta parte demuestran que tanto en el apartado de procesamiento como de análisis de corpus existen herramientas a nuestra disposición elaboradas por grupos investigadores de distintas universidades, algunas de ellas con un valor añadido, esto es, que se distribuyen libremente con fines investigadores. Nuestro siguiente objetivo fue presentar una propuesta metodológica para la elaboración de DICTUM, para la cual determinamos como principio metodológico fundamental el del hábitat natural. Asimismo, y siguiendo la estela de lo ya realizado en proyectos anteriores, pensamos que el corpus ha de contener etiquetas lingüísticas que expliciten la categoría gramatical de cada ítem que lo compone, pues de este modo el corpus se puede explotar de modo que la información que se obtiene es mucho más sustanciosa que cuando este recurso se encuentra en bruto, pues con este último formato el extractor actúa aplicando únicamente criterios estadísticos. En la extracción, por su parte, es aconsejable utilizar un extractor híbrido, por razones de calidad de los resultados; si bien es cierto que, la carencia de un extractor de

este tipo, podrá equilibrarse al haber etiquetado el corpus. Finalmente, adoptamos un modelo de entrada terminográfica combinatoria bilingüe, que presentamos a modo de boceto. Para albergar los lemas, las combinaciones y la distinta información asociada del modelo, pensamos que una base de datos relacional resulta un recurso electrónico eficaz y flexible para organizar y registrar la información lingüística, conceptual, pragmática y administrativa.

7. Bibliografía

ANTHONY, L. (2004). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. En *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*. 7-13.

ARAYA R. y VIVALDI, J. (2004). Mercedes: a term-context highlighter. En *IV International Conference on Language Resources and Evaluation. LREC 2004*. 445-448.

ARNTZ, Rainer. 1993. Terminological Equivalence and Translation. En Sonneveld, B. y Loening, K.L. (eds.). *Terminology: Applications in Interdisciplinary Communication*. Amsterdam/ Filadelfia: John Benjamins. 5-19.

BANERJEE, S., & PEDERSEN, T. (2003). The Design, Implementation, and Use of the Ngram Statistics Package. En *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. México.

BENSON, M., BENON, E. e ILSON, R. (1986). *The BBI Dictionary of English Word Combinations*. Amsterdam/Philadelphia: John Benjamins [ed. rev. 1997].

BOLSHAKOW, I. A., & GELBUKH, A. (2002). Word Combinations as an Important Part of Modern Electronic Dictionaries. *Procesamiento del Lenguaje Natural*, (29), 47-54.

BOSQUE, I. (dir.) (2006): *Diccionario combinatorio práctico del español* contemporáneo. Madrid: Ediciones SM.

CABRÉ, M.T. (1999). *La terminología: representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

CABRÉ, M.T., LORENTE, M. y ESTOPÀ, R. (1996). Terminología y Fraseología. En *Actas del V Simposio Iberoamericano de Terminología*, México: RITERM. 67-84.

CABRÉ M.T., ESTOPÀ R. y VIVALDI J. (2001). Automatic Term Detection: a review of current systems. D. Bourigault, C. Jacquemin, M.-C. L'Homme (eds.). *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins.

CRYSTAL, D. (1985). *A Dictionary of Linguistics and Phonetics*. Oxford-New York: OUP.

DAGAN, I., CHURCH, K. W., y GALE, W. A. (1993). Robust Bilingual Word Alignment for Machine Aided Translation. *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*. 1-8.

ESTOPÀ, R. (1999). *Extracció de terminologia: elementos per a la construcció d'un SEACUSE*, Tesis Doctoral. Barcelona: Institut Universitari de Lingüística Aplicada, 1999. En línea: <http://

www.tdx.cesca.es/>. Consultado el 8 de septiembre de 2009.

FRIEDBICHLER, M., y FRIEDBICHLER, I. (2009). The promise of 'KWiC-Web'. *The Journal of the European Medical Writers Association*, *18*(1), 62-65.

GAMBIER, Y. (1991). Travail et vocabulaire spécialisés: prolégomènes à une socio-terminologie. *Meta*, 36. 8-15.

GLÄSER, R. (1994/5). Relations between Phraseology and Terminology with Special Reference to English. En *Alfa: Actes de langue française et de linguistique*, vol.7/8.41-60.

GREAVES, C. (2009). *ConcGram 1.0: a phraseological search engine*. Amsterdam: John Benjamins.

HASS, M. R. (1967): What belongs in a bilingual dictionary? F. Householder &. S. Saporta (eds.). *Problems in Lexicography*. Bloomington: Indiana University. Press. 45-50.

HEID, U. y FREIBOTT, G. (1991). Collocations dans une base de données terminologique et lexicale. *Meta*, 36 (1), 77-91.

GONZÁLEZ REY, M. I. (2002). Contribución a una reflexión sobre las colocaciones. En Veiga, A., González Pereira, M y Souto Gómez, M. (eds.). *Léxico y gramática*. Lugo: Tris Tram. 155-171.

IULA (2006). Metodología de Trabajo en Terminología [en línea]. En *Grup IulaTerm. Tallers online de Terminologia*. Barcelona: IULA. Universidad Pompeu Fabra.http://www.iula.upf.edu/iulonlca.htm

KILGARRIFF, A., RYCHLY, P. S., & TUGWELL, D. (2004). The Sketch Engine. In G. Williams & S. Vessier. En *Proceedings of Euralex 2004*. Lorient: Université de Bretagne Sud.

KOIKE, K. (2001): *Colocaciones léxicas en el español actual: estudio formal y léxico-semántico*. Alcalá de Henares: Universidad de Alcalá y Takushoku University.

L'HOMME, M. (2000). Understanding specialized lexical combinations. *Terminology*, 6(1), 89-110.

L'HOMME, M. (2009). A methodology for describing collocations in a specialised dictionary. En prensa. John Benjamins Publishing Company.

LORENTE, M. (2002). Terminología y fraseología especializada: del léxico a la sintaxis. En GUERRERO RAMOS, G. y PÉREZ LAGOS, M.F. (eds.). *Panorama actual de la terminología*, Granada: Editorial Comares. 159-179.

LORENTE, M., BEVILACQUA, C. y ESTOPÀ, R. (1998). El análisis de la fraseología especializada mediante elementos de la lingüística actual. En Correia, M. (ed.) *Terminologia, desenvolvimiento é identidade nacional. VI Simposio Iberoamericano de Terminología.* La Habana, noviembre de 1998, Lisboa: ILTEC, Colibri. 647-666.

MARTIN, W. (1992). Remarks on Collocations in Sublanguages. *Terminologie et Traduction* 2/3, Bruselas: Commission des Communautés Européennes, Service de Traduction. 157-164.

MEYER, I. y MACKINTOSH, K. (1996). Refining the terminographer's concept-analysis methods: How can phraseology help? *Terminology*, 3(1). 1-26.

MOLINA PLAZA, S. (2006). The making of a bilingual dictionary of phraseological units English-Spanish/ Spanish-English with corpora examples. *Panace@*, 7(23), 99-105.

RUIZ GURILLO, L. (2002). Compuestos, colocaciones, locuciones: intento de deli-

mitación. En VEIGA, A., GONZÁLEZ PEREIRA, M y SOUTO GÓMEZ, M. (eds.). *Léxico y gramática*. Lugo: Tris Tram. 327-339.

SCOTT, M. (2008). *WordSmith Tools version 5*, Liverpool: Lexical Analysis Software.

SERETAN, V. (2008). *Collocation Extraction Based on Syntactic Parsing*. Tesis doctoral. Université de Genève, Faculté des lettres. Département de linguistique.

SERETAN, V., NERIMA, L., & WEHRLI, E. (2004). A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora. In G. Williams & S. Vessier, *Proceedings of Euralex 2004* (págs. 755-766). Lorient, Francia.

SMADJA, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, *19*(1), 143-177.

TODIRAȘCU, A., HEID, U., ȘTEFĂNESCU, D., TUFIȘ, D., GLEDHILL, C., WELLER, M., et al. (2008). Vers un dictionnaire de collocations multilingue. *Cahier de Linguistique*, *33*(1), 161-186.

VARGAS SIERRA, C. (2008). La sistematización terminográfica: una propuesta metodológica para la elaboración de diccionarios traductológicos. En *Actas del X Simposio Iberoamericano de Terminología*, Montevideo, Uruguay.

VERDAGUER, I., POCH, A., LASO, N. J., & GIMÉNEZ, E. (2008). Scie-Lex. A lexical database of collocations in scientific English for Spanish scientist. En *XIII Congreso Internacional Euralex*. Barcelona.

VIDAL, V. (2007). Consideraciones en torno a la descripción terminográfica de la combinatoria léxica especializada: aspectos macroestructurales. En Lorente, M; Estopà, R.; Freixa, J.; Martí, J. y Tebé, C. (eds.). *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví*. Barcelona: IULA, págs. 473-488.

WALKER, D. G. (1993). Translation Problems as They Occur in Everyday Practice. En *Terminology and Knowledge Engineering*. 221-224.