VOLUMEN MONOGRÁFICO (2005), 67-84

LAS COMBINACIONES LÉXICAS EN EL INGLÉS CIENTÍFICO: PRESENTACIÓN DE UNA BASE DE DATOS

ISABEL VERDAGUER CLAVERA
Universitat de Barcelona
MARÍA JUAN GARAU
Universitat de les Illes Balears

RESUMEN. *El objetivo de este artículo es presentar una base de datos de combinaciones léxicas en el inglés científico, que se está elaborando en tres universidades españolas (Barcelona, Illes Balears y León) y está primordialmente destinada a la comunidad científica hispanohablante. La constatación de la falta de diccionarios especializados que ofrecieran información contextual sobre los patrones gramaticales y combinatorios en registros específicos determinó la puesta en marcha del proyecto. Esta base de datos está centrada en el estudio de un corpus de textos escritos de aproximadamente cuatro millones de palabras de las áreas de la biología, bioquímica y biomedicina, y proporciona la información morfológica, semántica, sintáctica y combinatoria necesaria para el uso correcto y preciso de cada término en el discurso científico. En el presente trabajo se describen los pasos seguidos en la elaboración de la base de datos y se expone el estudio de caso de una de sus entradas.*

PALABRAS CLAVE: *Base de datos, combinaciones léxicas, inglés científico, corpus escrito.*

ABSTRACT. *The aim of this paper is to present a lexical database of English collocations used in scientific language, which is being built in three Spanish universities (Barcelona, Illes Balears and León) and is mainly intended for the Spanish-speaking scientific community. The shortage of specialized dictionaries providing contextual information on the grammatical and collocational patterns in specific registers prompted the onset of this project. Our database is based on the analysis of a corpus of written texts in the areas of biology, biochemistry, and biomedicine, and provides the grammatical, semantic, and collocational information necessary for the correct and precise use of each term in scientific discourse. The paper describes the steps followed in the creation of the data base and it includes the case study of one of its entries.*

KEYWORDS: *Database, collocations, scientific English, written corpus.*

1. Introducción[1]

En los textos científicos en lengua inglesa, al igual que ocurre sin duda en otras lenguas, aparecen de forma recurrente una serie de patrones léxicos que los miembros de la comunidad científica internacional no hablantes de dicho idioma como primera lengua, los científicos de habla hispana en nuestro caso, deben aprender no sólo a comprender sino también a utilizar adecuadamente para poder comunicarse de manera eficiente y precisa. Es justamente el uso de estas combinaciones léxicas lo que confiere fluidez y naturalidad al discurso. Hay que tener en cuenta, además, que estas combinaciones son específicas para cada lengua, por lo que los hablantes no nativos tienden a usar combinaciones inapropiadas en L2, especialmente en usos deslexicalizados, esto es, con escaso significado en el contexto en el que aparecen como por ejemplo *make* en *make a start* (véase Altenberg y Granger 2001).

Por todo ello nuestro trabajo actual se centra en la elaboración de una base de datos de aquellas combinaciones léxicas que están presentes habitualmente en la prosa científica en lengua inglesa. Más concretamente, y como primer estadio en la elaboración de nuestra base de datos, hemos creado un corpus a partir de textos especializados en los ámbitos de la biología, la bioquímica y la biomedicina escritos en inglés. La base de datos nos permitirá generar de manera prácticamente automática un diccionario combinatorio del inglés científico que proporcione información sobre las combinaciones más usuales en esta lengua de especialidad. Esta herramienta está destinada a la comunidad científica de habla española –especialmente bioquímicos, médicos, biólogos y farmacéuticos– así como a traductores técnicos y a profesores o estudiantes de inglés para fines específicos.

El presente artículo se ha organizado del modo siguiente. En primer lugar nos detendremos en el lugar destacado que el estudio de las combinaciones léxicas recurrentes ocupa en el campo de la lexicografía. El segundo apartado hace referencia a los pasos seguidos a la hora de confeccionar y analizar el corpus a partir del cual se crea la base de datos, mientras que el tercer apartado se dedica a la información contenida en ella. A continuación se incluye un estudio de caso de una entrada que ejemplifica lo expuesto en el apartado anterior. Por último, se exponen algunas consideraciones finales sobre el trabajo realizado.

1.1. La importancia de las combinaciones léxicas

El proceso de construcción del discurso podría describirse someramente como un proceso de combinación de palabras. No obstante, a la hora de generar dichas combinaciones, los hablantes de una lengua no son enteramente libres. De hecho, su libertad se ve constreñida por dos tipos de limitaciones: sintácticas y léxicas. Los patrones sintácticos generales de una lengua son los que determinan qué combinaciones sintácticas son posibles en ella. Por su parte, las restricciones léxicas se manifiestan en el hecho de que en una lengua determinada algunas palabras suelen aparecer juntas, mientras que otras

no. Sinclair (1991) enfatiza la relevancia del segundo tipo de restricción al contraponer lo que se ha dado en llamar *open-choice principle* al *idiom principle*. De acuerdo con el primer principio, las restricciones que operan en la construcción del significado son básicamente de tipo gramatical con la consecuencia de que prácticamente cualquier palabra puede llenar un hueco determinado (*slot-and-filler model*). El *idiom principle*, en cambio, reconoce la importancia de las colocaciones típicas (*collocations*) a las que el hablante recurre asiduamente, entre otras razones, para economizar esfuerzos o para afrontar las exigencias de la conversación en tiempo real. Este último principio es mucho más dominante de lo que se había reconocido anteriormente y contribuye al menos tanto como la gramática a explicar la producción textual.

Algunas de las principales características de las colocaciones típicas incluyen la dificultad en ciertos casos para establecer cuáles son los límites de la colocación, la variación léxica interna, la variación sintáctica interna, o la co-ocurrencia con determinadas opciones gramaticales o determinados entornos semánticos (Sinclair 1991: 111-112). Igualmente, a raíz del análisis de corpus lingüísticos, se ha observado que existe un gran número de palabras deslexicalizadas, así como de palabras cuya primera definición en una entrada de diccionario resulta no representar ni con mucho el significado más habitual de la misma (p. e. *back* es mucho más frecuente en sus usos adverbiales que para referirse a la espalda humana). Todo ello ha llevado a los lexicógrafos a hablar de usos de palabras más que de significados.

De todo lo anterior se desprende que los diccionarios deben ir más allá de la definición de palabras como unidades aisladas. Es esencial que también indiquen al usuario cómo las palabras se combinan entre sí para formar frases. Las combinaciones léxicas presentan diferentes grados de cohesión que nos permiten clasificarlas en diversos grupos. Benson, Benson y Ilson (1986) distinguen entre combinaciones libres, expresiones idiomáticas (*idioms*), colocaciones típicas, combinaciones de transición, y palabras compuestas, si bien las fronteras entre unas y otras no son siempre fáciles de trazar. La mayor parte de combinaciones léxicas entrarían dentro de la categoría de combinaciones libres y como su propio nombre indica son aquellas que permiten una mayor libertad combinatoria. Así por ejemplo, el nombre *murder* puede aparecer ligado a una significativa cantidad de verbos (*analyze, boast of, condemn, describe, discuss, disregard, examine, film, forget, investigate, mention, recall, record, remember, report, romanticize, study,* etc., *a murder*) (Benson, Benson y Ilson 1986: 252). Por su parte, las expresiones idiomáticas constituyen un grupo de expresiones relativamente fijas cuyo significado no refleja los significados de sus partes constituyentes. Así pues el significado de *kick a dust* (armar un lío) no es deducible a partir de sus componentes, *kick* y *dust*. Dentro de este bloque también pueden situarse los refranes o proverbios (p. e. *Time and tide wait for no man*), aunque a menudo tengan un sentido más literal. A medio camino entre las expresiones idiomáticas y las combinaciones libres se sitúan las colocaciones típicas. Es el caso, por ejemplo, de *commit murder*, una combinación que se registra mucho más frecuentemente que las anteriormente reseñadas con relación al nombre *murder* y que además apenas tiene variación posible (el único sinónimo en este

contexto sería *perpetrate*). Estas características hacen que Benson, Benson y Ilson la denominen combinación fija, combinación recurrente o colocación. Dichos autores se refieren a las combinaciones a caballo entre las colocaciones típicas y las expresiones idiomáticas como combinaciones de transición. Se trata de expresiones más fijas que las colocaciones típicas pero cuyo significado, a diferencia de las expresiones idiomáticas, es deducible a partir de sus componentes como, por ejemplo, *on the spur of the moment*. Por último, las palabras compuestas se forman mayoritariamente en inglés a partir de adjetivo + nombre o nombre + nombre. Señalaríamos a modo de ejemplo el compuesto *red herring* (pista falsa), a veces clasificado como expresión idiomática, que aparece como entrada independiente en la sexta edición del *Oxford Advanced Learner's Dictionary* (Hornby 2000). El mayor problema para los lexicógrafos ha sido sin duda el tratamiento de las colocaciones típicas, puesto que son más difíciles de identificar que las expresiones idiomáticas o incluso que las palabras compuestas. Por esta razón, su inclusión en diccionarios ha sido un tanto errática.

A través de la elaboración de nuestra base de datos, hemos identificado los diferentes tipos de combinaciones clasificadas por Benson, Benson e Ilson (1986), si bien nos hemos centrado en aquéllas que son más abundantes y recurrentes en nuestro corpus, incluyendo tanto las colocaciones típicas, como las combinaciones libres más frecuentes, además, naturalmente, de las expresiones idiomáticas y combinaciones de transición. Conviene apuntar que las combinaciones léxicas no se han clasificado en la base de datos de acuerdo a la categorización expuesta, a no ser en el apartado de notas del investigador, ya que consideramos que dicha información no es relevante para los futuros usuarios del diccionario combinatorio.

Desde una perspectiva didáctica, en los últimos tiempos se ha destacado asimismo el papel preeminente que las combinaciones léxicas juegan en la adquisición y buen manejo de una lengua segunda o extranjera (Richards y Rogers 2001). Mientras que la influyente teoría de Chomsky de adquisición de lenguas ponía de relieve la capacidad de los hablantes de crear frases jamás producidas anteriormente, el llamado Enfoque Léxico (Willis 1990; Nattinger y DeCarrico 1992; Lewis 1993, 1997, 2000; Woolard 2000) mantiene que tan sólo una ínfima parte de las producciones lingüísticas son enteramente novedosas y que, por el contrario, las unidades léxicas de más de una palabra funcionan como fragmentos (*chunks*) o patrones memorizados que contribuyen en un porcentaje muy elevado a la producción discursiva. Las colocaciones típicas ocupan un lugar preferente dentro de este enfoque. Saber utilizarlas adecuadamente se considera esencial para que el hablante no nativo consiga transmitir su mensaje de manera eficiente y correcta.

En los últimos años han ido apareciendo una serie de diccionarios combinatorios en el mercado editorial que avalan la importancia del estudio lexicográfico de las colocaciones típicas. En el caso de la lengua inglesa, entre dichos diccionarios se encuentran *A Dictionary of English Collocations: Based on the Brown Corpus* (Kjellmer 1994), *The BBI Dictionary of English Word Combinations* (Benson, Benson e Ilson 1997), *The LTP Dictionary of Selected Collocations* (Hill y Lewis 1998), *Collins*

Cobuild English Words in Use: a Dictionary of Collocations (1999) o el recientemente publicado *Oxford Collocations Dictionary for Students of English* (2004). También existen algunos diccionarios que examinan la equivalencia de diferentes combinaciones léxicas en dos lenguas, como el diccionario de Benson y Benson (1992) *Russian-English Dictionary of Verbal Collocations*.

La expansion del inglés como *lingua franca* en el ámbito científico también ha contribuido a la difusión de estudios sobre el inglés académico (Swales 1990; Alcaraz 2000; Fernández y Gil 2000; García 2000; López 2001; Flowerdew 2002, entre muchos otros) y más específicamente sobre fraseología y colocaciones en este registro (Howarth 1996; Tercedor 1999; Williams 1999; Gledhill 2000; Oakey 2002). Sin embargo, después de una exploración de los diccionarios a disposición de los científicos de habla no inglesa, constatamos la falta de obras que proporcionen información acerca del uso y las combinaciones de palabras generales en el inglés científico. Existen buenos diccionarios tanto monolingües –en los que se encuentra información enciclopédica sobre terminología especializada, prácticamente restringida a sustantivos–, como bilingües o multilingües –proporcionando el equivalente en otras lenguas–, pero la mayoría de ellos no incluyen información sobre palabras no especializadas (véase Norman 2002; L'Homme 2003). Exceptuando un diccionario dedicado a un ámbito muy especializado, *Parasitic Plant Dictionary*[2], no existe ningún otro diccionario del que tengamos conocimiento que ayude al hablante no nativo del inglés a usar las combinaciones adecuadas de palabras en el discurso científico.

2. Pasos en la creación de la base de datos

En este apartado se describen los pasos seguidos a la hora de crear la base de datos: la selección del corpus, su procesamiento con la ayuda de la aplicación informática *WordSmith Tools*, y finalmente el análisis de las líneas de concordancia para establecer los diferentes usos y combinaciones léxicas de cada palabra.

2.1. Selección y descripción del corpus

El primer paso en la creación de la base de datos es la selección del corpus. En este caso, se ha partido de un corpus del inglés escrito previamente creado por nuestro equipo investigador que constaba de un millón de palabras aproximadamente y se centraba en el análisis del lenguaje real utilizado en el área de la bioquímica (Verdaguer y Juan 1998-2000). El corpus actual se ha ampliado hasta incluir unos cuatro millones de palabras. Si bien dicha cifra no sería suficiente para un corpus del lenguaje general, ya que éstos suelen constar de entre diez y veinte millones de palabras, consideramos que es adecuada para estudiar un sublenguaje específico como se pretende.

El corpus que hemos confeccionado abarca los ámbitos no sólo de la bioquímica, sino también de la biología y la biomedicina. Se basa en artículos de revistas científicas

de reconocido prestigio, disponibles en formato electrónico, entre las que destacan *Biochemical Journal, Genes and Development, British Medical Journal,* y *The Journal of Cell Biology*.

2.2. Estudio del corpus con WordSmith Tools

Una vez obtenido el corpus, se necesita un programa de gestión del mismo que automatice ciertas tareas lexicográficas. Para la gestión del presente corpus se ha utilizado el programa *WordSmith Tools*, gracias al cual hemos obtenido en primer lugar el listado de términos presentes en el corpus, de entre los cuales se han seleccionado los que presentaban una frecuencia superior a cinco. Conviene señalar en este punto que no dedicamos nuestros esfuerzos exclusivamente a estudiar la combinatoria de vocablos técnicos, sino también, y muy especialmente, de palabras del inglés general que presentan unas combinaciones específicas en el inglés científico.

WordSmith Tools además de listas de palabras y concordancias (ver Figura 1 a continuación), proporciona al usuario *collocates* y *clusters*, es decir, listas con las combinaciones más frecuentes. En el caso de los *collocates*, según se observa en la Figura 2, se indica la frecuencia con que determinadas palabras aparecen a la izquierda o a la derecha de la palabra estudiada (señalada con un asterisco). Así, por ejemplo, puede verse que la palabra *principle* aparece inmediatamente precedida por *in* en 46 ocasiones (ello se simboliza con la expresión L1, esto es, primera palabra a la izquierda) y seguida por *of* en 18 casos (dato que se simboliza con la expresión R1, es decir, primera palabra a la derecha). Por lo que se refiere a los *clusters*, en la Figura 3 puede observarse como el programa nos proporciona los conjuntos de tres palabras que aparecen repetidamente en el texto, sin signos de puntuación que las separen, ordenados de mayor a menor frecuencia de aparición. Así puede verse que la combinación *the principle of* es la más habitual en este caso. Ambas funciones, *collocates* y *clusters*, dan al investigador una buena idea de las características combinatorias de la palabra objeto de estudio.

N	Concordance
1	position or methylation explanations because in principle almost any sequence character could af
2	mined to assess which is the more reliable. This principle also applies when two different molecula
3	t. This has become known as the Haldane-Muller principle and implies that the mutational load de
4	er than its condemnation? Can any agreement on principle be found? Can respect for different view
5	accounted for by the parameter K, which can in principle be estimated and compared to U. The
6	gene flux through the E. colichromosome could in principle belong any number of distinct bacterial
7	cent common (female) ancestor. The coalescent principle can be used to estimate the number of
8	atrix metalloproteinases (MMPs) are among the principle classes of proteinases that facilitate th
9	o eyes, are simultaneously active.(75) The same principle could apply to converging visual and au
10	ody plan. Indeed, this selection pressure may in principle explain both the initial assembly and su
11	duces aggregation. While this mechanism can in principle explain how different scrapie strains ca
12	tal Health, Glasgow). Contributors: DMG was the principle grant applicant and provided clinical inpu
13	2). This result demonstrates the counter-intuitive principle (implicit in Metabolic Control Analy- sis
14	hemselves, their sacrifice on behalf of an abstract principle is without their consent. However, as n
15	specific combinations. A simple example of this principle is found in the specification of cell-type
16	stigated in cdc28-1N and cks1 mutants. Clb2, the principle mitotic cyclin of S. cerevisiae (Grandin a
17	poor or inconsistent management strategies, are principle obstacles to successful zoo breeding [C
18	treat analyses that do not comply with the basic principle of analysing all randomised subjects as
19	weak acid (ammonium/propionate). The general principle of the technique is that cells exposed t
20	as the approach used in most trials, violates the principle of intention to treat and leads to bias unl
21	edure for separating biological particles using the principle of a counterstreaming centrifuge. A con
22	ver them, but then, to my mind, there is only one principle of development, and that is that The be
23	tion of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Bioch
24	es of suppression (Figure 1) were isolated. The principle of isolation of the phx mutations is desc
25	l loss rate in subsequent years.(62) Although the principle repair mechanism within the endotheliu
26	sion, any microtubule tethered at an end could in principle retain the capacity to flow at this associ
27	parated by less than 11 Å; -amine of lysine is the principle side chain target for this NHS-ester (25).
28	uals, and 3) dependence on a single, fundamental principle such as sentience or self-awareness in
29	hich is several-fold above its IC &l and therefore in principle sufficient to exert an autoinhibitory effect
30	concept of an "entelechy:ö an internal perfecting principle that he believed characterized all living
31	ulations are less variable, however, suggesting in principle that it would be easier to reject a model
32	esults demonstrate the currently counter-intuitive principle that it is the fraction of the genome unde
33	e axonal tracts or brain structures than others, a principle that is echoed for knockouts of related
34	for conformational change (Figure 5B). Thus, in principle the MUG enzymes appear to provide a
35	r over the whole genome, U. Haldane applied this principle to estimate the mutation load in Droso

Figura 1. Ejemplo de concordancias.

N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	THE	81	47	34	8	10	4	8	17	0	1	15	6	9	3
2	PRINCIPLE	77	0	0	0	0	0	0	0	77	0	0	0	0	0
3	IN	70	64	6	12	1	2	3	46	0	0	1	2	1	2
4	OF	52	14	38	2	4	4	3	1	0	18	3	4	6	7
5	PRINCIPLES	50	0	4	0	0	0	0	0	46	0	1	1	1	1
6	TO	26	9	17	0	3	3	3	0	0	4	3	3	4	3
7	AND	25	8	17	2	2	1	2	1	0	9	0	4	1	3
8	THAT	22	9	13	1	1	1	6	0	0	4	3	1	2	3
9	BE	20	4	16	2	2	0	0	0	0	2	9	1	3	1
10	IS	17	8	9	3	0	3	2	0	0	2	1	2	2	2
11	A	15	8	7	3	2	1	0	2	0	0	3	2	1	1
12	THIS	15	8	7	1	1	1	0	5	0	0	3	1	1	2
13	COULD	14	6	8	0	0	0	6	0	0	1	0	2	4	1
14	ARE	11	7	4	1	1	2	2	1	0	3	0	0	1	0
15	GENERAL	10	8	2	0	0	1	1	6	0	0	0	0	1	1
16	CAN	9	4	5	0	1	0	3	0	0	1	0	3	1	0
17	IT	8	3	5	0	2	1	0	0	0	0	4	0	0	1
18	ON	8	7	1	1	2	2	1	1	0	1	0	0	0	0
19	OR	8	4	4	3	0	0	0	1	0	0	1	1	1	1
20	BY	7	4	3	1	2	0	1	0	0	0	0	1	2	0
21	AN	6	5	1	2	1	1	1	0	0	0	0	0	1	0
22	AS	6	3	3	0	1	0	1	1	0	0	1	2	0	0
23	BIOLOGY	6	1	5	0	0	1	0	0	0	0	0	3	2	0
24	CELL	6	4	2	0	0	4	0	0	0	0	2	0	0	0
25	FOR	6	2	4	1	0	0	0	1	0	0	0	1	2	1
26	ANY	5	2	3	0	0	1	1	0	0	0	2	1	0	0
27	AT	5	3	2	1	1	1	0	0	0	0	0	0	1	1
28	BASIC	5	5	0	0	0	0	0	5	0	0	0	0	0	0
29	FROM	5	1	4	0	0	0	1	0	0	0	1	0	2	1
30	GENE	5	2	3	1	0	1	0	0	0	0	1	0	0	2
31	NOT	5	3	2	1	2	0	0	0	0	0	1	0	1	0
32	PRACTICE	5	0	5	0	0	0	0	0	0	0	2	3	0	0
33	SHOULD	5	4	1	1	0	0	3	0	0	0	1	0	0	0
34	SUCH	5	2	3	1	0	0	0	1	0	0	3	0	0	0
35	WITH	5	4	1	1	1	2	0	0	0	0	1	0	0	0

Figura 2. Ejemplo de *collocates.*

N	Cluster	Freq.
1	the principle of	4
2	the principles of	4
3	and clinical practice	3
4	basic principles and	3
5	of invasion biology	3
6	principles and clinical	3
7	principles of invasion	3
8	sickle cell disease	3
9	that the principles	3

Figura 3: Ejemplo de *clusters.*

2.3. Análisis de concordancias

Nuestro punto de partida para el estudio léxico del sublenguaje que nos ocupa son los textos especializados. A partir de ellos, y con la ayuda de las diferentes herramientas de *WordSmith* antes mencionadas (ver 2.2.), hemos procedido al análisis de cada una de las unidades lingüísticas que aparecen en la base de datos. Para ello se ha comenzado por buscar las concordancias de cada término, esto es, las listas de palabras en su contexto de uso. *WordSmith* nos proporciona por defecto la línea en la que aparece la palabra objeto de estudio, pero dicho contexto es susceptible de ser ampliado hasta incluir las líneas anteriores y posteriores necesarias para llevar a término dicha labor.

Pensando en nuestros destinatarios, se ha procurado incluir en la base de datos información exhaustiva sobre las combinaciones sintácticas e idiomáticas del inglés científico de modo que éstos puedan utilizarla para interpretar y generar correctamente nuevos textos científicos. Cuando abordamos la descripción de un nuevo término, en primer lugar establecemos si tiene o no diferentes significados. En segundo lugar, estudiamos los diferentes patrones léxico-sintácticos asociados con el uso de la palabra en cuestión en cada uno de sus posibles significados. Por último, examinamos las posibles colocaciones que de manera recurrente aparecen en relación con dichos patrones. A continuación se desgrana la información incluida en la base de datos.

3. Información incluida en la base de datos léxica

Nuestro objetivo es ayudar al hablante no nativo de inglés –y muy especialmente al de lengua española– a codificar su mensaje usando no solamente una estructura gramatical correcta, sino también las combinaciones léxicas más apropiadas. Por tanto, y con el fin de almacenar y manejar la información que consideramos necesaria para el usuario, hemos desarrollado una base de datos encaminada a facilitar su producción lingüística. Mientras que los diccionarios técnicos normalmente están dirigidos a un uso pasivo, reflejado en el tipo de información que ofrecen, el objetivo de este diccionario es facilitar, como hemos indicado, la producción lingüística del usuario. Nuestra base de datos proporciona la información necesaria para poder utilizar en el sublenguaje científico los términos incluidos con precisión.

Como puede apreciarse en la Figura 4, cada entrada principal viene representada por una ventana a partir de la cual se puede acceder a subventanas que dan cabida a las posibles ramificaciones de la misma (pertenencia a diferentes categorías gramaticales; distintos sentidos dentro de una misma categoría; diversos patrones combinatorios léxicos y sintácticos dentro de cada sentido).

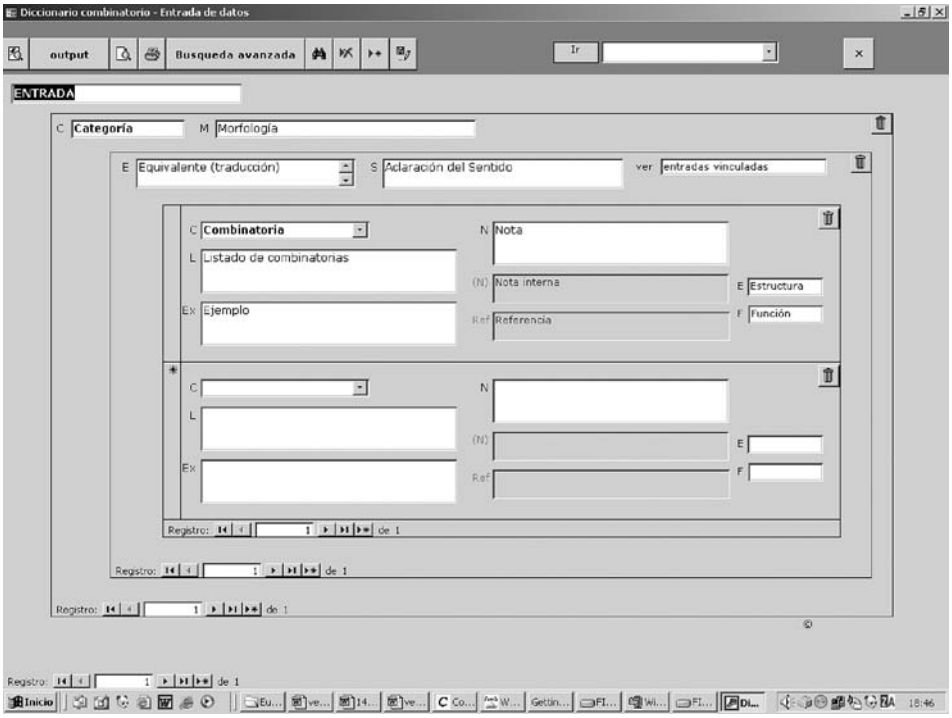

Figura 4. Pantalla de tratamiento de entradas.

Cada entrada proporciona la siguiente información:

Primer bloque

- **C**: Categoría gramatical (Nombre (N), Adjetivo (Adj), Verbo (V), Adverbio (Adv), Preposición (Prep)). Éste es el primer parámetro a tener en cuenta, puesto que el sentido, las características morfológicas y formas flexivas, el comportamiento sintáctico y también a veces la pronunciación vienen determinados por la categoría gramatical. Si un término puede pertenecer a más de una categoría, se abrirá un registro para cada una, donde la información dependerá de la categoría en cuestión.
- **M**: Morfología. Variantes morfológicas (N: singular / plural; V: forma base / 3ª persona singular / pasado / participio / *-ing*). Hemos decidido incluir esta información sobre las formas nominales y verbales, tanto si son regulares como irregulares, para facilitar la producción lingüística del usuario.

Segundo bloque

- **E**: Equivalencia terminológica en español.

- **S**: Aclaración del sentido. Si una palabra es polisémica ya bien en inglés o en español, hay que aclarar el sentido en que se utiliza en cada ocasión por medio de una glosa o términos sinónimos. Así por ejemplo los múltiples sentidos del verbo *take* equivalen frecuentemente a *tomar* en español, pero hay que distinguir sus numerosas acepciones (tomar medicamentos, tomar imágenes, etc.).
- **Ver**: Entradas vinculadas. Las referencias cruzadas a otras entradas pueden ser útiles en varios casos, como por ejemplo cuando las palabras están relacionadas morfológicamente o semánticamente (sinónimos, antónimos, etc.).

Tercer bloque

- **C**: Construcciones gramaticales en las que puede aparecer cada sentido. La interacción del significado y la valencia de un término –es decir los elementos que este término requiere– es crucial, ya que en muchas ocasiones los distintos sentidos de un término se expresan por medio de distintos patrones sintácticos. Esta información es clave para la correcta construcción de la oración, especialmente cuando la entrada es un verbo.
- **L**: Combinaciones léxicas más frecuentes. Aquí incluimos la lista de los términos que en nuestro corpus aparecen más recurrentemente combinados con el término tratado. Están organizados por bloques semánticos y, dentro de cada bloque, están ordenados alfabéticamente.
- **Ex**: Ejemplos de uso real. Los ejemplos seleccionados ilustran y completan la información proporcionada en la entrada. Estos ejemplos están inspirados en el corpus, aunque no son transcripciones exactas de las frases encontradas, que a menudo son de gran complejidad.
- **N**: Notas aclaratorias para destacar usos especiales o ayudar al usuario a utilizar correctamente un término, que se añaden cuando el investigador lo cree conveniente.

Esta base de datos duplica de manera infinita cada uno de los tres bloques, tanto el que recoge la información sobre la categoría gramatical del término, como el que informa sobre su equivalencia terminológica, y el que proporciona información sintáctica y combinatoria. Dicha prestación permite la generación de entradas de estructura compleja que contengan subentradas para cada una de las diferentes categorías gramaticales a las cuales pertenezca el lema en tratamiento, al igual que para cada uno de sus sentidos y de sus distintas complementaciones, que, a su vez, determinan las diversas combinaciones léxicas.

Además de esta información, que es la que se proporcionará en el *output* destinado al usuario, en nuestra base de datos se han previsto unos campos adicionales: las notas no destinadas al usuario, sino al equipo investigador; la referencia de los ejemplos; y unos campos con información funcional y estructural más detallada de interés para los estudios teóricos del equipo.

4. Estudio de principle

La unidad léxica seleccionada para ilustrar el estudio lingüístico que llevamos a cabo y la información que se almacena en nuestra base de datos es *PRINCIPLE.* Éste es un término que no está restringido al lenguaje especializado, pero que naturalmente aparece con cierta frecuencia en los textos científicos.

PRINCIPLE es una palabra polisémica que en el lenguaje general se utiliza a menudo en el sentido de norma de conducta personal o de un conjunto de normas morales. En el *Collins English Dictionary and Thesaurus* (2000), donde se presenta el sentido más frecuente de una palabra en primer lugar, éstas son las dos primeras de siete acepciones (más dos frases hechas). *WordNet*, por otra parte, identifica seis sentidos distintos. Sin embargo, en un lenguaje especializado, los términos poseen unos significados específicos y más restringidos (Pearson 1998).

A continuación mostramos el contenido y el funcionamiento de la base de datos con el estudio que hemos realizado del término *PRINCIPLE* en el lenguaje científico:

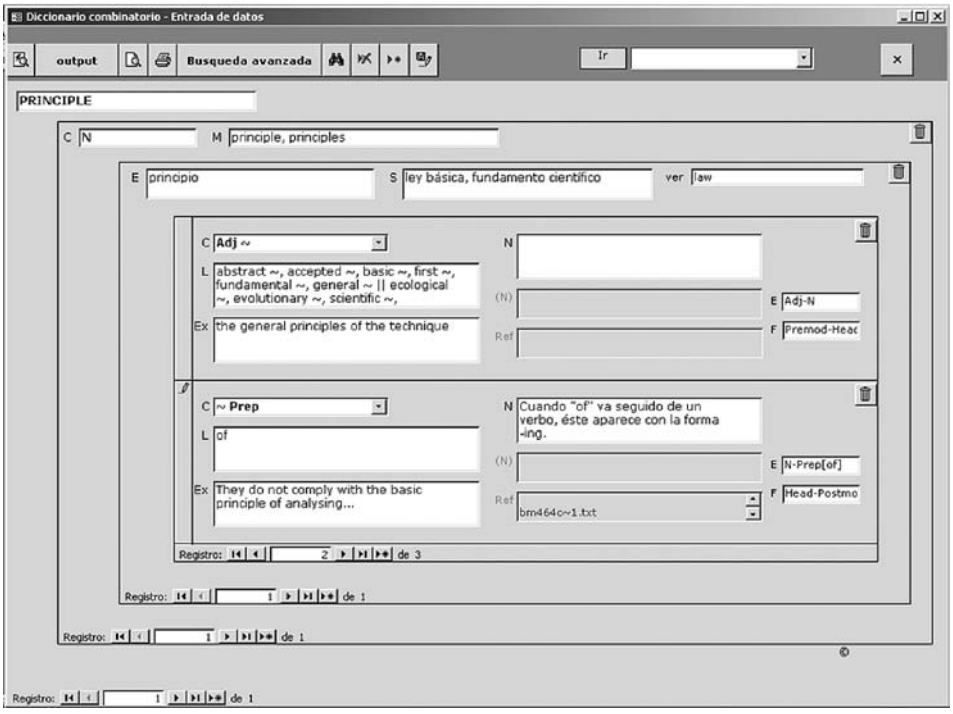

Figura 5. *PRINCIPLE* (1).

En el primer bloque de la Figura 5 se indica que *PRINCIPLE* es un sustantivo[3]. Como es un término contable, en la información morfológica se incluye el plural del término. El siguiente bloque muestra la información semántica correspondiente a la primera acepción del término. Su equivalente terminológico en español es "principio" y como ésta es una palabra polisémica se aclara a continuación el sentido en que se utiliza (ley básica, fundamento científico). Aunque en un análisis semántico muy afinado podrían considerarse dos subsentidos, uno más general (ley básica) y otro más específico (fundamento científico), en nuestra base de datos hemos optado por tratarlos como una única acepción debido a la dificultad de establecer una clara distinción semántica en muchos de los contextos en los que aparece y a la similitud de sus patrones combinatorios en este sublenguaje. El próximo campo refiere al usuario a un término de significado parecido.

Finalmente, en el último bloque se incorporan los patrones combinatorios gramaticales y léxicos de este sentido. Con el fin de evitar al máximo etiquetas lingüísticas especializadas, hemos optado por especificar la argumentación de cada uno de los usos verbales de la manera más sencilla posible y emplear un sistema de codificación formal en lugar de funcional. En primer lugar se indica que *PRINCIPLE* puede ir premodificado por un adjetivo (Adj ~) y a continuación se incluye una lista de los adjetivos más frecuentes que pueden acompañar este término, ordenados alfabéticamente y por bloques semánticos (*abstract, accepted, basic, first, fundamental, general* || *ecological, evolutionary, scientific, toxicological*), amén de un ejemplo. En este caso, no es necesaria ninguna nota aclaratoria para el usuario ni tampoco indicar la referencia del ejemplo, que es muy general. En los campos de Estructura y Función, que no aparecerán en el *output* por no ser de interés para los usuarios no lingüistas a quienes va destinada esta base de datos, se ha especificado Adj-N (Estructura) y *Premodifier-Head* (Función).

El siguiente registro muestra *PRINCIPLE* seguido de la preposición *of*. Aquí interesa destacar esta posibilidad y, naturalmente, no hay listado de las palabras que pueden seguir dicha preposición, pues sería infinito. Nos limitamos a señalar un ejemplo y en una nota para el usuario indicamos que, si *of* va seguido de un verbo, éste tiene que ir en la forma *–ing*. En este caso, la estructura es N–Prep (*of*) y la función *Head-Postmodifier*.

Hasta este momento hemos visto las posibles combinaciones de *PRINCIPLE* como núcleo de un sintagma nominal, pero también hay que incluir sus posibles combinatorias dentro de la oración, que se muestran en la Figura 6.

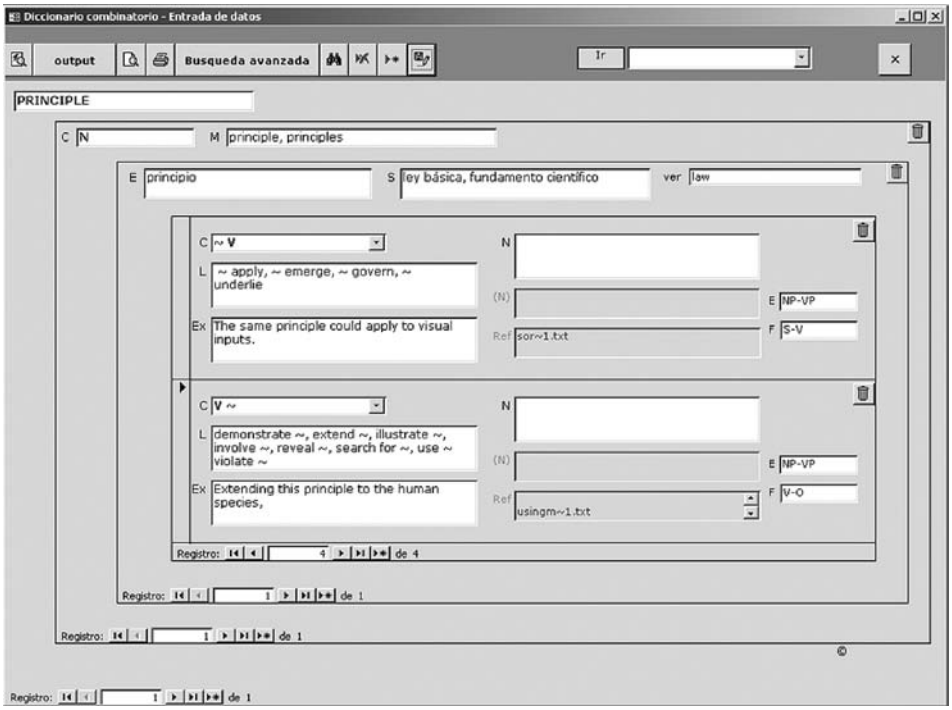

Figura 6. *PRINCIPLE* (2).

En este sentido *PRINCIPLE* puede ser utilizado como sujeto y, por lo tanto, anteceder a determinados verbos (~ V), que se especifican en el siguiente campo (*apply, emerge, govern, underlie*). La estructura, pues, sería NP–VP y la función S–V. También puede aparecer como objeto directo de verbos transitivos como *demonstrate, extend, illustrate, involve, reveal, search for, use* o *violate* y, por tanto, sigue al verbo, lo cual se expresa de la forma (V ~). En los apartados de estructura y función destinados al equipo investigador las notaciones serían NP–VP y V–O respectivamente.

La siguiente pantalla muestra como *PRINCIPLE* también puede ir seguido de una *that*-clause apositiva, lo cual se especifica en el siguiente registro (ver Figura 7), con un ejemplo ilustrando esta posibilidad. La estructura de esta combinación sería N–*that*-clause, y la función, *Head–Postmodifier*.

La palabra tratada aparece también frecuentemente en la frase idiomática *in principle* (en principio). Aquí se ha añadido una nota para el equipo investigador, indicando su frecuente aparición con verbos modales.

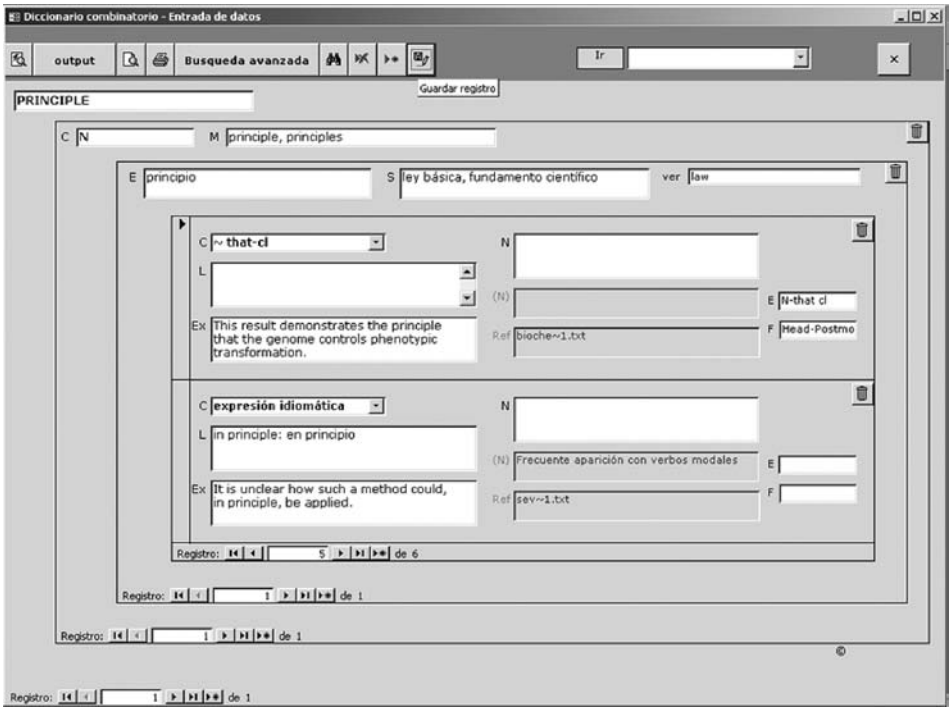

Figura 7. *PRINCIPLE* (3).

Finalmente, *PRINCIPLE* puede tener otro significado, que equivale igualmente en español a "principio", pero con el sentido más específico "constituyente que proporciona las características propias de una sustancia" (ver Figura 8). Por lo tanto, se activará de nuevo el segundo bloque y se añadirá la información correspondiente. En este sentido, *PRINCIPLE* se combina con adjetivos, en especial *active*, pudiéndose considerar esta combinación un compuesto. La estructura es Adj–N, y la función, *Premodifier–Head*.

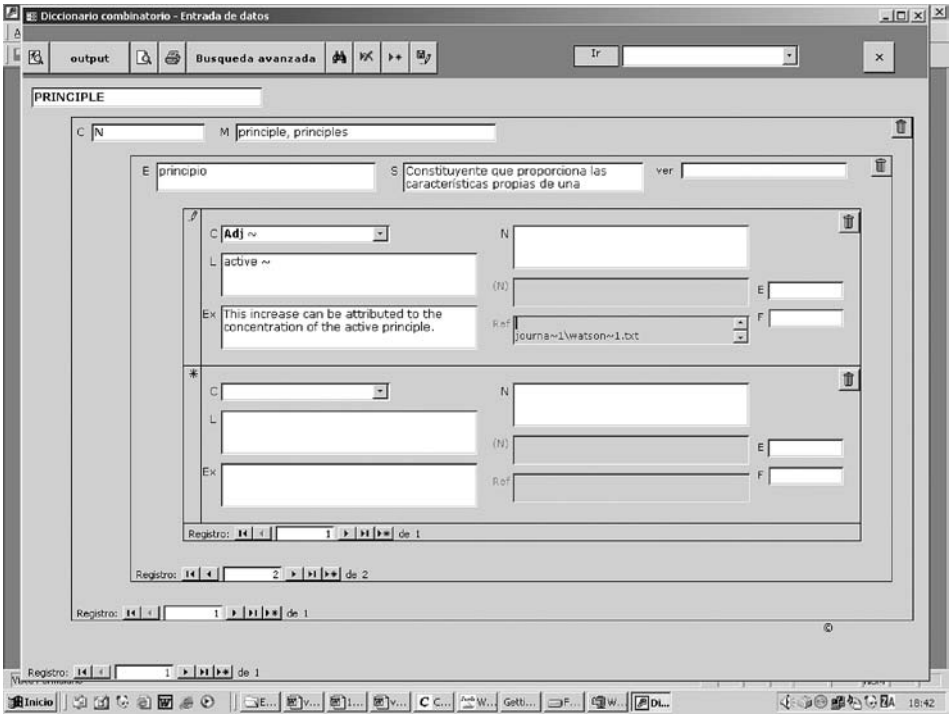

Figura 8. *PRINCIPLE* (4).

5. Conclusiones

La entrada que hemos ofrecido ha mostrado cómo nuestra base de datos proporciona la información necesaria (morfológica, semántica, sintáctica y combinatoria) que permite al usuario el uso correcto y preciso de cada término en el discurso científico. La necesidad de una obra de referencia de estas características quedó patente en el Congreso Internacional EURALEX 2004, que tuvo lugar recientemente en la Universidad de Bretaña Sur (Verdaguer y González 2004). Creemos, por lo tanto, que tendrá una muy buena acogida entre la comunidad científica de habla española. En estos momentos, la base de datos está en proceso de elaboración y hemos previsto que en la fase previa a su introducción sea objeto de unas pruebas piloto de utilización y evaluación por parte de futuros usuarios.

6. Notas

1. Este proyecto, referencia BFF2001-2988, está financiado por el Ministerio de Ciencia y Tecnología y FEDER.

2. Geoffrey Williams está trabajando en un diccionario altamente especializado basado en un corpus sobre plantas parásitas, *Parasitic Plant Dictionary* (http://perso.wanadoo.fr/geoffrey.williams/), en el que también incluye términos no especializados.
3. Aunque en esta base de datos parece que falta la pronunciación de los términos, un aspecto muy a tener en cuenta en un diccionario dirigido a la producción lingüística, tenemos proyectado incluir la pronunciación audible de los términos, evitando la transcripción fonética, que es de difícil interpretación si no se tiene formación lingüística.

7. Referencias bibliográficas

Alcaraz, E. 2000. *El inglés profesional y académico.* Madrid: Alianza Editorial.

Altenberg, B. y S. Granger. 2001. "The grammatical and lexical patterning of MAKE in native and non-native student writing". *Applied Linguistics* 22 (2): 173-195.

Benson, M. y E. Benson. 1992. *Russian-English Dictionary of Verbal Collocations*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Benson, M., E. Benson y R. Ilson. 1986. *Lexicographic Description of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Benson, M., E. Benson y R. Ilson. 1997. *The BBI Dictionary of English Word Combinations*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Collins English Dictionary and Thesaurus. 2000. Aylesbury: HarperCollins Publishers.

Collins Cobuild English Words in Use: a Dictionary of Collocations. 1999. New York: HarperCollins Publishers.

Flowerdew, J. 2002. *Academic Discourse*. Edinburgh: Pearson Education Limited.

Fernández, F. y L. Gil. 2000. *Enlaces oracionales y organización retórica del discurso científico en inglés y en español*. Valencia: Universitat de València.

García, M. P. 2000. *English for Specific Purposes: Discourse Analysis and Course Design.* Universidad del País Vasco.

Gledhill, C. J. 2000. *Collocations in Science Writing*. Tübingen: Gunter Narr Verlag.

Hill, E. y M. Lewis, eds. 1998. *The LTP Dictionary of Selected Collocations*. London: Language Teaching Publications.

Hornby, A. S. 2000. *Oxford Advanced Learner's Dictionary of Current English*. Oxford: Oxford University Press.

Howarth, P. H. 1996. *Phraseology in English Academic Writing*. Tübingen: Max Niemeyer Verlag.

Kjellmer, G. 1994. *A Dictionary of English Collocations: Based on the Brown Corpus*. Oxford: Oxford University Press.

Lewis, M. 1993. *The Lexical Approach.* London: Language Teaching Publications.

Lewis, M. 1997. *Implementing the Lexical Approach.* London: Language Teaching Publications.

Lewis, M. 2000. "Learning in the Lexical Approach". *Teaching Collocation: Further Developments in the Lexical Approach*. Ed. M. Lewis. London: Language Teaching Publications. 155-184.

L'Homme, M. C. 2003. "Verbs and Verbal Derivatives. A Model for Specialized Lexicography". *International Journal of Lexicography* 16.4: 403-422.

López, B. 2001. *Estudio descriptivo comparado inglés/español de la representación del conocimiento en los "abstracts" de las ciencias de la salud.* Tesis doctoral. Universidad de Valladolid. Documento disponible en http://www.cervantesvirtual.com.

Nattinger, J. R. y J. S. DeCarrico. 1992. *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Norman, G. 2002. "Description and Prescription in Dictionaries of Scientific Terms". *International Journal of Lexicography* 15.4: 259-276.

Oakey, D. 2002. "Formulaic Language in English Academic Writing". *Using Corpora to Explore Linguistic Variation*. Eds. Reppen, R. *et al.* Amsterdam/Philadelphia: John Benjamins Publishing Company. 111-129.

Oxford Collocations Dictionary for Students of English. 2004. Oxford: Oxford University Press.

Pearson, J. 1998. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Richards, J. C. y T. S. Rodgers. 2001. *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press. Chapter 12.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Swales, J. M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Tercedor, M. 1999. *La fraseología en el lenguaje biomédico: análisis desde las necesidades del traductor*. Documento disponible en http://www.elies.rediris.es.

Verdaguer, I y E. González. 2004. "A lexical database of collocations in scientific English: Preliminary considerations". *Euralex 2004 Proceedings.* Lorient. Vol 3: 929-934.

Verdaguer, I. y M. Juan. 1998-2000. "Generación de un diccionario especializado combinatorio bilingüe". *Anuari de Filologia*. Vols. XXI-XXII. Secció A. Número 9. 69-78.

Williams, G. 1999. *Les Réseux Collocationnels dans la Construction et l'Exploitation d'un Corpus dans le Cadre d'une Communauté de Discours Scientifique.* Tesis Doctoral. Documento disponible en http://perso.wanadoo.fr/geoffrey.williams/.

Willis, J. D. 1990. *The Lexical Syllabus*. London: Collins COBUILD.

Woolard, G. 2000. "Collocation-encouraging learner independence". *Teaching Collocation: Further Developments in the Lexical Approach.* Ed. M. Lewis. London: Language Teaching Publications. 28-46.

WordNet. Disponible en http://www.cogsci.princeton.edu.