Secondary bibliography

Alberdi, Cristina. 1999. "La violencia del género". *El País* (18-02-1999).
Bergvall, Victoria. 1999. "Toward a comprehensive theory of language and gender." *Language and Society* 28: 273-293.
Garre, Marianne. 1999. *Human Rights in Translation: Legal concepts in different languages*. Copenhagen: Copenhagen Business School Press.
Lázaro Carreter, Fernando. 2000. "El dardo en la palabra: Vísperas navideñas". *El País* (3-12-2000).
Molina Foix, Vicente. 1999. "El género epiceno". *El País* (9-03-1999)
Oakley, Ann. 1972. *Sex, Gender, and Society*. Aldershot, UK: Arena.
Valdecantos, Camilo. 1999. "Sexo, sólo sexo". *El País* (7-03-1999)
Wagner, Emma *et al.* 2001. *Translating for the European Institutions*. Manchester: St. Jerome.

CHAPTER TWENTY-THREE

ACCESO A LA INFORMACIÓN TERMINOLÓGICA EN INTERNET: TÉCNICAS PARA TRADUCTORES

VERÓNICA PASTOR AND AMPARO ALCINA
TECNOLETTRA, UNIVERSITAT JAUME I

1. Introducción

En los últimos años, los traductores acceden a nuevos recursos, como los corpus e Internet, para buscar la terminología que necesitan. En este sentido, resulta lógico, como reclaman investigadores y profesores, que sea necesario incorporar a estas nuevas herramientas en los programas de formación de futuros traductores y estudiantes de lenguas. Algunos recursos de búsqueda en Internet no incluyen manuales de uso, o cada manual explica de forma distinta las funciones que pueden utilizarse en cada recurso. Existen pocas clasificaciones que traten de sistematizar todas las técnicas de búsqueda que pueden utilizarse en recursos similares de una forma coherente.

En investigaciones anteriores, el grupo de investigación TecnoLeTTra, en el marco de los proyectos ONTODIC y ONTODIC II[1], hemos desarrollado clasificaciones de técnicas de búsqueda que reúnen las principales técnicas de búsqueda en recursos diferentes de manera estándar basándonos en tres parámetros: consulta, instrumento y resultado. Estas clasificaciones nos han permitido organizar las técnicas de búsqueda en diccionarios electrónicos y corpus utilizando los mismos parámetros de clasificación. En este estudio aplicamos este sistema de clasificación a Internet.

2. Recursos analizados

En esta investigación hemos efectuado un análisis de herramientas que permiten la búsqueda de terminología en Internet. Hemos analizado algunos buscadores, como Google, Altavista y AlltheWeb. Asimismo, hemos descrito herramientas de búsqueda de Internet como corpus, WebCorp (Renouf, Kehoe y Banerjee 2007), WebCONC (Hüning, Freie Universität Berlin), KWiCFinder (Fletcher 2007a) y Web Concordancer beta (Fletcher 2007b). Por último, hemos repasado herramientas que permiten la compilación y análisis de textos de Internet, en concreto, WordSmith 4, Corpógrafo (Sarmento, Maia y Santos 2004) y TerminoWeb (Barrière 2009).

3. Clasificación de técnicas de búsqueda en Internet

Las técnicas de búsqueda son opciones que el usuario puede utilizar en un recurso para obtener un determinado resultado. Nuestra clasificación de técnicas de búsqueda está dividida en función de los tres elementos que hemos visto que intervienen en una búsqueda: la consulta, el instrumento y el resultado. La *consulta* es la palabra o expresión que el usuario introduce en la interfaz de un recurso. El *instrumento* es el recurso o parte de un recurso en el que se busca la palabra o expresión de búsqueda. El *resultado* de la búsqueda es el elemento al que se accede cuando se consulta un recurso.

En la siguiente ilustración aparece representada nuestra clasificación de técnicas de búsqueda en la imagen de un pescador que introduce en el agua un cebo para pescar un pez. Introduce como cebo una consulta, la palabra exacta *play*, por ejemplo en la interfaz de un corpus monolingüe inglés (que sería en este caso el instrumento) y obtiene como resultado de la pesca una lista de concordancias de la palabra *play*, entre las que se incluyen expresiones como *play the piano, play football* o *play the role of.*

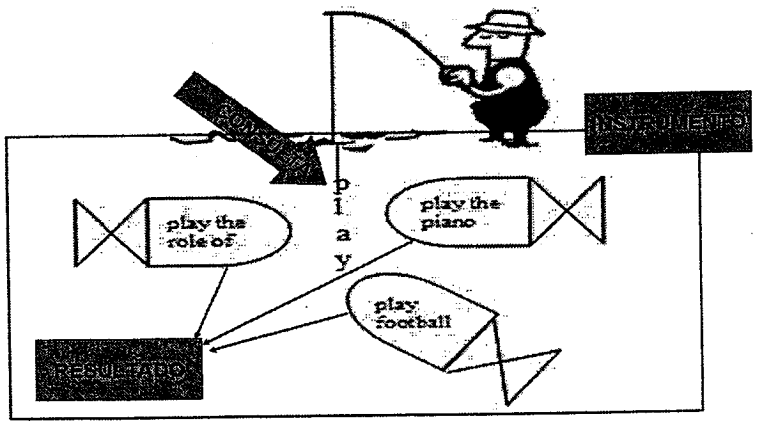


Fig. 23-1. Representación de los tres elementos de nuestra clasificación de técnicas de búsqueda (Pastor y Alcina 2009)

3.1. La consulta

El primer elemento de nuestra clasificación es la consulta. Una **consulta** puede ser una **expresión léxica**. Las expresiones léxicas que introducimos en Internet pueden ser formas exactas o truncadas o secuencias exactas o truncadas.

Una **forma exacta** es una palabra completa Las formas exactas pueden introducirse en cualquiera de las herramientas analizadas de Internet. Por ejemplo en KWiCFinder hacemos una búsqueda de las palabras exactas *llueve*, *cántaros* y *raudales* en páginas de Internet en español (Fig. 23-2). En el resultado vemos concordancias en las que aparecen dichas palabras de búsqueda. Accedemos a expresiones como *llueve sobre mojado, llover a cántaros, llueve a raudales,* etc.

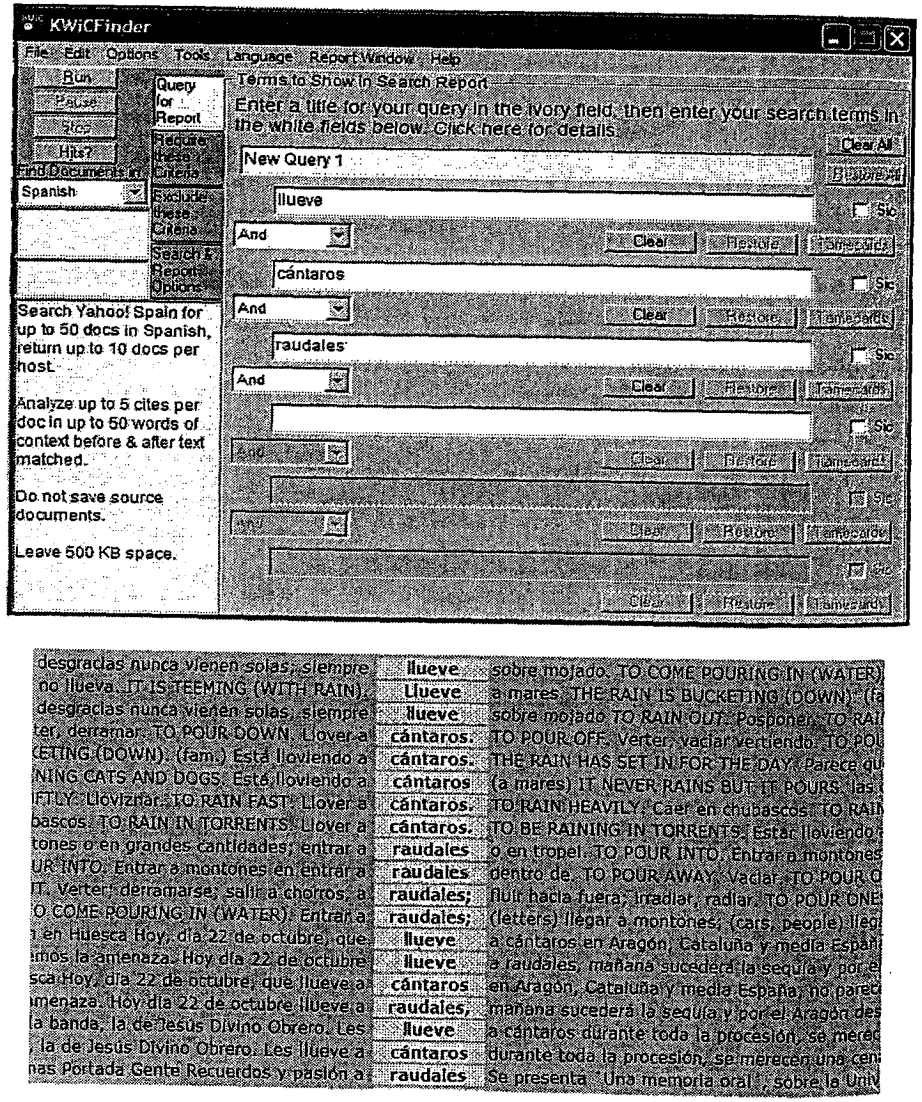


Fig. 23-2. Búsqueda de formas exactas en KWiCFinder

Una **forma truncada** es una palabra incompleta. La parte de la palabra que se omite, se sustituye con un carácter comodín. Puede ser útil si se quieren buscar todas las palabras que contengan, empiecen o terminen con una secuencia de caracteres. Los buscadores no admiten el uso de comodines para truncar palabras. Algunas herramientas de análisis de la

Web como corpus tampoco porque dependen de los buscadores para recuperar los textos. WebCorp, KWiCFinder y WordSmith 4 admiten el uso de comodines en las búsquedas. WebCONC permite el uso de expresiones regulares cuando el usuario especifica su propio corpus de búsqueda.

Una **secuencia exacta** es una combinación de formas que se busca en el recurso de la misma forma en que se ha introducido. Es posible introducir secuencias exactas en los buscadores, por medio de su introducción entre comillas, y también en el resto de herramientas analizadas en Internet.

Una **secuencia truncada** es una combinación de formas en la que se omite alguna palabra y se sustituye con un comodín. Puede ser útil si se quiere buscar una expresión y solo se conoce alguna de las palabras que la componen. Pueden introducirse secuencias truncadas en los buscadores utilizando comodines dentro de las comillas. También en WebCorp, si introducimos la secuencia truncada *no * ojo,* se recuperan concordancias con las secuencias: *no tiene ojo, no pegas ojo, no quitar ojo, no pierde ojo,* etc. WebCONC permite introducir secuencias truncadas sólo cuando el traductor especifica el corpus de búsqueda. KWiCFinder permite utilizar comodines para buscar secuencias truncadas sólo cuando las páginas ya se han recuperado de Internet, para buscar nueva terminología dentro de los resultados. WordSmith 4 permite la introducción de secuencias truncadas.

La consulta también puede ser un **número**. Puede ser útil para localizar palabras que suelen aparecer en un contexto próximo a algún número significativo. En todas las herramientas de búsqueda en Internet pueden introducirse números en las búsquedas, pero los resultados no son tan significativos como con los corpus. Lo vemos con un ejemplo. Al introducir en la interfaz de Bwananet[2] el número 640 para buscarlo en un corpus de la informática en español, localizamos palabras como *píxel,* porque una medida que recupera el corpus es *640x480 píxels*; también se localiza *memoria RAM*, porque otra medida que se recupera es *640 Kb de memoria RAM*. Todas estas expresiones pertenecen al ámbito de la informática. Sin embargo, si realizamos esta misma búsqueda, del número 640, en un buscador como Google, recuperamos expresiones de áreas muy variadas como *Real Decreto* 640*/2009*, *Teka TR*-640, *Motos KTM LC4 E* 640, etc. Por otro lado, Google tiene una opción de búsqueda que permite buscar intervalos de números, añadiendo dos puntos entre las dos cifras del intervalo, por ejemplo, la búsqueda de *125.512 Mb*, recuperará las medidas *128 Mb*, *256 Mb*, *512 Mb*, *etc.*

La **combinación discontinua de expresiones** consiste en introducir en el corpus un elemento determinado y establecer la distancia en la que debe

aparecer también un segundo elemento. De las herramientas de búsqueda en Internet que hemos analizado, WordSmith 4 permite la introducción de combinaciones discontinuas de expresiones, especificando el número de posiciones a izquierda y derecha entre cada expresión. El buscador Altavista permite el uso del operador NEAR para buscar expresiones cercanas. Por ejemplo, si introducimos la expresión en Altavista *Cytomegalovirus NEAR "is a"*, en los resultados encontramos contextos definitorios del término *Cytomegalovirus*, como por ejemplo, ***Human cytomegalovirus is a member of the herpes virus family.***

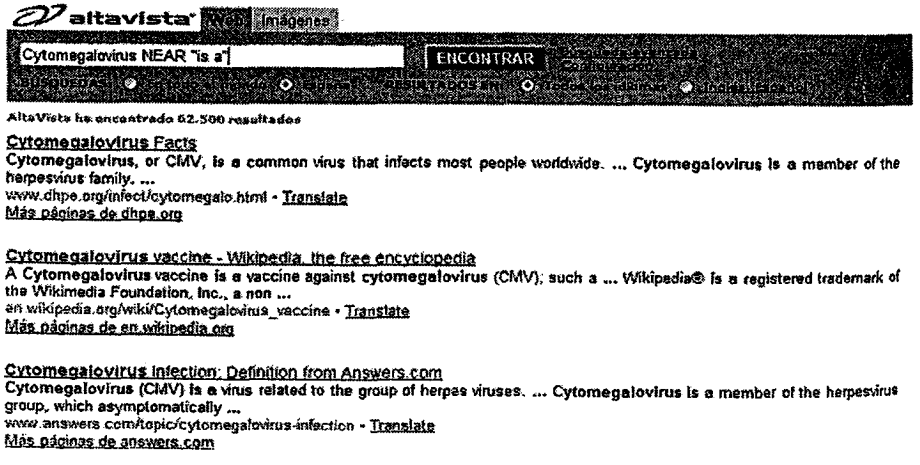


Fig. 23-3. Búsqueda de una combinación discontinua de expresiones en Altavista

El programa TerminoWeb contiene la función *Term pair exploration* en la que el usuario selecciona dos términos de búsqueda, y el programa recupera los contextos en los que aparecen en una posición cercana.

Los **filtros** añaden una condición de búsqueda a la consulta introducida. En Internet pueden utilizarse filtros de inclusión y exclusión de palabras, filtros de restricción de idioma, de área geográfica, área temática, fecha, formato de archivo, etc. Por ejemplo podemos restringir palabras con el uso de los operadores AND, OR, NOT, en los buscadores. También en Web Concordancer beta pueden utilizarse este tipo de filtros, como se ve en la siguiente imagen. Introducimos la palabra *pasta* y restringimos por idioma a páginas en español e incluimos en la búsqueda las palabras *construcción* y *edificios* pero excluimos *comida, cocina, macarrones*. El resultado son contextos en los que *pasta* aparece en páginas del ámbito de la construcción, pero no de cocina. Vemos un

ejemplo de concordancias extraídas de una página sobre tipos de porcelana.

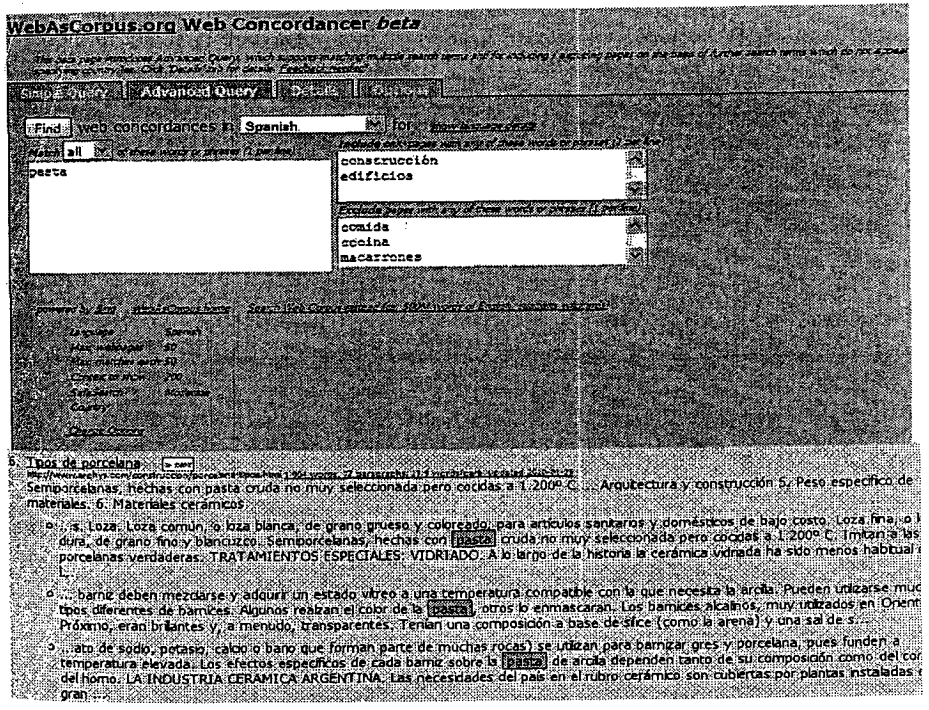


Fig. 23-4. Filtros de inclusión y exclusión de palabras en Web Concordancer beta

3.2. El instrumento

Hemos clasificado las herramientas de búsqueda de terminología en Internet principalmente en tres tipos de instrumento. Dependiendo del instrumento, hemos visto que las consultas y los resultados pueden variar. El instrumento más habitual para la búsqueda en Internet son los **buscadores**, como por ejemplo Google, Altavista y AlltheWeb.

Asimismo, hemos encontrado un nuevo tipo de instrumento desarrollado para la búsqueda en Internet como si fuera un corpus (**Web as Corpus**). Algunos ejemplos son WebCorp, WebCONC, KWiCFinder, o Web Concordancer beta. La diferencia principal de estas herramientas y los buscadores radica en el modo en que se presentan los resultados de la búsqueda. Los buscadores recuperan páginas web. Debajo de cada página recuperada aparece un resumen de la página con las palabras de búsqueda

resaltadas. Para observar el uso de las palabras de búsqueda en contexto, el usuario tiene que acceder a cada página ofrecida por el buscador. En cambio, las herramientas de búsqueda en Internet como corpus se sirven de los buscadores de Internet para la recuperación de las páginas analizadas, pero los resultados de la búsqueda se visualizan a modo de concordancias KWIC (todos los contextos aparecen de manera seguida y con las palabras de búsqueda centradas y destacadas, con contexto a izquierda y derecha). Algunos de estos recursos ofrecen también otras opciones, como generación de colocaciones o de listas de palabras.

Por último, hemos visto **herramientas para la compilación y análisis de textos de Internet**, como WordSmith 4, Corpógrafo y TerminoWeb. Estas herramientas facilitan la descarga de corpus virtuales desde Internet para su posterior análisis.

3.3. El resultado

El tercer elemento de nuestra clasificación, el **resultado**, puede ser una lista de contextos o concordancias, una lista de palabras, una lista de sinónimos, una lista de colocaciones, una lista de agrupaciones de palabras y una lista de imágenes.

Las **concordancias** son contextos en los que aparece la consulta que hemos introducido. Dependiendo de en qué instrumento de Internet se busque, la presentación de estos contextos puede variar. Realizamos una búsqueda de expresiones muy similares en instrumentos diferentes. En el buscador Google buscamos la secuencia truncada "llueve a *" y recuperamos un listado de páginas web en las que aparecen expresiones como "nunca llueve a gusto de todos" o "llueve a cántaros". Las palabras de búsqueda aparecen destacadas en el resumen de las páginas recuperadas, pero para acceder a los contextos completos, el usuario tiene que entrar en cada página recuperada. Sin embargo, una búsqueda similar de las palabras *llueve, cántaros,* y *raudales* en la herramienta de análisis de la web como corpus KWiCFinder, recupera los resultados a modo de concordancias KWIC, donde las palabras de búsqueda aparecen dentro de su contexto y destacadas en el centro, lo que facilita el análisis lingüístico de los resultados.

CHARLY MORLOCK: Nunca llueve a gusto de todos
3 Mar 2010 ... Nunca llueve a gusto de todos, que gran verdad: si hace pocas semanas comentaba lo bien que viene el agua al campo y las reservas ...
www.charlymorlock.com/.../nunca-llueve-gusto-de-todos.html - En caché

Diario de Jerez - Nunca llueve a gusto de todos
8 Mar 2010 ... Nunca llueve a gusto de todos. . ... Diario de Jerez; Jerez; Jerez; Nunca llueve a gusto de todos. Nunca llueve a gusto de todos ...
www.diariodejerez.es/.../llueve/.../todos.html - hace 17 horas - En caché

La Nueva España - Diario Independiente de Asturias - Occidente ...
9 Mar 2010 ... ¿Y si llueve a cántaros? CELSA DÍAZ ALONSO Otra vez. Otro nuevo retraso anunciado en la lenta autovía de nuestros dolores. ...
www.lne.es/occidente/2010/03/09/llueve.../883677.html - hace 6 horas

llueve a cántaros / llover a cántaros - WordReference Forums - [Traducir esta página]
12 entradas - 9 autores - Última entrada: 11 Dic 2009
llueve a cántaros / llover a cántaros Spanish-English Vocabulary / Vocabulario Español-Inglés.
forum.wordreference.com/showthread.php?t... - En caché - Similares

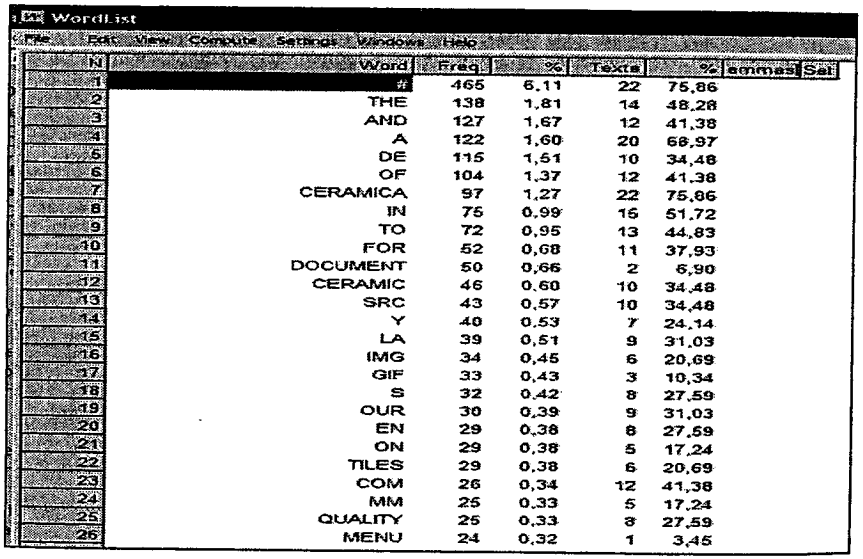
desgracias nunca vienen solas; siempre llueve	llueve
no llueve. IT IS TEEMING (WITH RAIN)	Llueve
desgracias nunca vienen solas), siempre llueve	llueve
ar, derramar. TO POUR DOWN. Llover a	cántaros.
CETING (DOWN). (fam.) Está lloviendo a	cántaros.
NING CATS AND DOGS. Está lloviendo a	cántaros
ELLY. Llovizna. TO RAIN FAST. Llover a	cántaros.
bascos. TO RAIN IN TORRENTS. Llover a	cántaros.
ones o en grandes cantidades; entrar a	raudales
OR INTO. Entrar a montones en entrar a	raudales
T. Verter; derramarse; salir a chorros; a	raudales;
O COME POURING IN (WATER). Entrar a	raudales;
en Huesca Hoy, día 22 de octubre, que	llueve
mos la amenaza. Hoy día 22 de octubre	llueve
sca Hoy, día 22 de octubre, que llueve a	cántaros
menaza. Hoy día 22 de octubre llueve a	raudales,
la banda, la de Jesús Divino Obrero. Les	llueve
Ida de Jesús Divino Obrero. Les llueve a	cántaros
nas Portada Gente Recuerdos y pasión a	raudales

sobre mojado. TO COME POURING IN (WATER)	
a mares. THE RAIN IS BUCKETING (DOWN). (fa	
sobre mojado TO RAIN OUT. Posponer. TO RAI	
TO POUR OFF. Verter; vaciar;vaciando. TO POU	
THE RAIN HAS SET IN FOR THE DAY. Parece qu	
(a mares) IT NEVER RAINS BUT IT POURS. las	
TO RAIN HEAVILY. Caer en chubascos. TO RAI	
TO BE RAINING IN TORRENTS. Estar lloviendo	
o en tropel. TO POUR INTO. Entrar a montones	
dentro de. TO POUR AWAY. Vaciar. TO POUR O	
fluir hacia fuera; irradiar; radiar. TO POUR ONE	
(letters) llegar a montones; (cars, people) lleg	
a cántaros en Aragón, Cataluña y media Espan	
a raudales, mañana sucederá la sequía y por el	
en Aragón, Cataluña y media España; no parec	
mañana sucederá la sequía y por el Aragón des	
a cántaros durante toda la procesión; se merec	
durante toda la procesión; se merecen una cen	
Se presenta: 'Una memoria oral' ; sobre la Univ	

Fig. 23-5. Comparación de resultados en Google y KWiCFinder

Las **listas de palabras** pueden ser de palabras más **frecuentes** del corpus y de palabras **clave**. Pueden ser útiles para tener una visión rápida de la terminología propia de un ámbito. Los buscadores no producen listas de palabras. WebCorp y WordSmith 4 generan listas de palabras. TerminoWeb tiene una función de extracción terminológica, que genera listas de palabras frecuentes de un corpus. WordSmith 4 es la única herramienta, de las analizadas, que genera listas de palabras clave. En la Fig. 23-6 vemos una lista de palabras frecuentes de páginas recuperadas por WordSmith 4 desde Internet sobre el campo cerámica. Entre los resultados vemos que las palabras más frecuentes son del lenguaje general. Pero dentro de la lista encontramos más palabras propias de la cerámica,

como *cerámica, ceramic, tile, brick, porcelain, mosaico, horno, terracota, arcilla, decoración, porcelana, pottery, engobe,* etc.



WordList

File Edit View Compute Settings Windows Help

N	Word	Freq	%	Texts	%	sgmts	Sel
1	#	465	6,11	22	75,86		
2	THE	138	1,81	14	48,28		
3	AND	127	1,67	12	41,38		
4	A	122	1,60	20	68,97		
5	DE	115	1,51	10	34,48		
6	OF	104	1,37	12	41,38		
7	CERAMICA	97	1,27	22	75,86		
8	IN	75	0,99	15	51,72		
9	TO	72	0,95	13	44,83		
10	FOR	52	0,68	11	37,93		
11	DOCUMENT	50	0,66	2	6,90		
12	CERAMIC	46	0,60	10	34,48		
13	SRC	43	0,57	10	34,48		
14	Y	40	0,53	7	24,14		
15	LA	39	0,51	9	31,03		
16	IMG	34	0,45	6	20,69		
17	GIF	33	0,43	3	10,34		
18	S	32	0,42	8	27,59		
19	OUR	30	0,39	9	31,03		
20	EN	29	0,38	8	27,59		
21	ON	29	0,38	5	17,24		
22	TILES	29	0,38	6	20,69		
23	COM	26	0,34	12	41,38		
24	MM	25	0,33	5	17,24		
25	QUALITY	25	0,33	8	27,59		
26	MENU	24	0,32	1	3,45		

Fig. 23-6. Lista de palabras en WordSmith 4

Las **listas de agrupaciones de palabras** son secuencias de dos o más palabras que aparecen con frecuencia en el corpus. Las agrupaciones de palabras ofrecen una visión rápida de cómo se combina la terminología de un ámbito. También permiten localizar una palabra a partir de otra palabra con la que se combina. De las herramientas analizadas, sólo WebCorp, Corpógrafo y WordSmith generan listas de agrupaciones de palabras.

Otro resultado son las **listas de sinónimos**. Algunos recursos han incorporado la función de búsquedas semánticas. Consiste en que se puede acceder a los sinónimos de la palabra introducida y realizar distintos tipos de búsqueda en relación con dichos sinónimos. Esta técnica de búsqueda puede ser útil para acceder a una palabra a partir de otras palabras que están relacionadas semánticamente. En las herramientas de búsqueda en Internet se están produciendo avances en la búsqueda de sinónimos y palabras relacionadas semánticamente. Google busca sinónimos de las palabras introducidas cuando aparecen precedidas de una vírgula ~. Esta técnica de búsqueda funciona parcialmente sólo con el inglés. Google Sets busca palabras relacionadas. Por ejemplo, si introducimos los términos *ratón* y *teclado*, se genera una lista de palabras del campo de la informática como *impresora, monitor, escáner, tarjeta de red, tarjeta de sonido*, etc.

TerminoWeb contiene una función de búsqueda de patrones semánticos que rodean a una palabra o palabras del corpus. Los patrones disponibles en TerminoWeb son de antonimia, causa, definición, función, hiperonimia, meronimia, similitud y sinonimia. El usuario puede definir nuevos tipos de relaciones e introducir patrones nuevos.

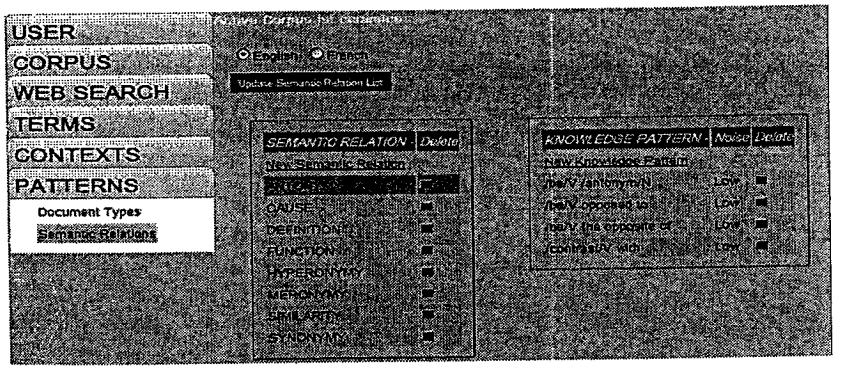


Fig. 23-7. Relaciones semánticas y patrones en TerminoWeb

En cuanto a las **listas de colocaciones**, una colocación es una palabra que aparece frecuentemente cercana a otra palabra. Las colocaciones permiten obtener una visión rápida de las palabras que aparecen en un contexto inmediato a otra palabra.

De las herramientas analizadas, las únicas que generan listas de colocaciones son WebCorp, KWiCFinder, TerminoWeb y WordSmith 4. En WebCorp, buscamos el término *hormigón* y generamos una lista de colocaciones: *de, impreso, en, hormigón, pulido, para, Empresa, armado, estructuras, pavimentos; Madrid, prefabricadas, del, madera, casas, a, desventajas, soleras* y *metálicas*.

Por último, algunas herramientas de búsqueda en Internet tienen como resultado una **imagen o una lista de imágenes**. Las imágenes pueden resultar útiles a los traductores para averiguar a qué hace referencia una palabra, cuando no es posible encontrar una definición. Todos los buscadores permiten buscar imágenes. Éste resultado no se encuentra en los corpus, pero sí en los diccionarios electrónicos. Por ejemplo, en Google podemos generar una lista de imágenes a partir de la introducción de la palabra *mosaico*.