CONFLUÊNCIAS - Revista de Tradução Científica e Técnica, N.º 4, Maio 2006: 84-107

EL PROCESO TERMINOGRÁFICO MULTILINGÜE CON *WORDSMITH TOOLS*

CHELO VARGAS SIERRA

Departamento de Filología Inglesa, Universidad de Alicante, España

Resumen:

Un corpus especializado sirve de base en la tarea terminológica para varias funciones, que se pueden reducir, fundamentalmente, a cuatro: a) *identificar los candidatos a término;* b) *proporcionar más datos sobre dichos posibles términos (combinatoria, derivados, relaciones con otros términos, etc.);* c) *ayudar a la compilación y elaboración de las definiciones;* d) *ofrecer un buen número de ejemplos contextuales. Sin embargo, el primer problema que se le puede plantear en la extracción de datos a partir de un corpus no etiquetado es cómo proceder ordenadamente para poder aislar los términos relevantes al ámbito objeto de estudio. El objetivo de este artículo es presentar, en primer lugar, una propuesta metodológica para analizar tal corpus, dividida en cuatro aproximaciones o fases; en segundo lugar, proporcionar una forma metódica de utilizar unos determinados instrumentos informáticos, que también serán descritos; y, en tercer lugar, de forma combinada con dichos instrumentos, mostrar una serie de estrategias para la detección y extracción semiautomática de terminología.*

Palabras clave:

Terminología; Extracción terminológica; Corpus; Programas de concordancias.

Resumo:

Um *corpus* especializado serve de base, na tarefa terminológica, para várias funções, que se podem reduzir, fundamentalmente, a quatro: *a)* identificar os candidatos a termo; *b)* proporcionar mais dados sobre os ditos possíveis termos (combinatória, derivados, relações com outros termos, etc.); *c)* ajudar à compilação e elaboração das definições; *d)* oferecer um bom número de exemplos contextuais. No entanto, o primeiro problema que se pode colocar na extracção de dados a partir de um *corpus* não etiquetado é como proceder

ordenadamente para poder isolar os termos relevantes ao âmbito objecto de estudo. O objectivo deste artigo é apresentar, em primeiro lugar, uma proposta metodológica para analisar um *corpus* com essas características, dividida em quatro aproximações ou fases; em segundo lugar, proporcionar uma forma metódica de utilizar determinadas ferramentas informáticas, que também serão descritas; e, em terceiro lugar, de forma combinada com as referidas ferramentas, mostrar diversas estratégias para a detecção e extracção semiautomática de terminologia.

Palavras-Chave:
Terminologia; Extracção terminológica; *Corpus*; Programas de concordâncias.

Abstract:
A special-purpose corpus serves, within terminology tasks, as a basis for performing several functions, which can be reduced, basically, to four: a) *to identify candidate terms;* b) *to provide more data about these possible terms (combinations, derivatives, relations with other terms, etc.);* c) *to help compile and elaborate definitions;* d) *to offer an important number of contextual examples. However, the first problem a terminologist may encounter when retrieving data from a raw corpus is how to proceed methodically so as to be able to isolate terms relevant to the subject-matter field. The objective of this article is, firstly, to present a four-step methodological proposal to analyze such a corpus; secondly, to provide a systematic way to use certain computer instruments, which will be also described; and, thirdly, in combination with these instruments, to show a series of strategies for the semiautomatic recognition and extraction of terminology.*

Keywords:
Terminology; Terminology extraction; Corpus; Concordancers.

1. INTRODUCCIÓN

La compilación de un corpus no es un fenómeno nuevo, sino que para determinadas disciplinas, como la lexicografía y la enseñanza de lenguas, se ha constituido desde épocas pasadas como una práctica común. A finales de los años setenta y principios de los ochenta se popularizan los ordenadores personales, situación que hizo posible el acceso de un mayor número de investigadores al procesamiento del lenguaje natural. En los años ochenta es cuando se extiende el uso de córpora junto con herramientas informáticas para su procesamiento y explotación. Este nuevo contexto tecnológico y social contribuyó de manera decisiva en el resurgimiento y fortalecimiento de la investigación lingüística basada en corpus. La vinculación del ordenador a la compilación y creación de grandes córpora ha hecho que hoy en día el término corpus contenga en sí la característica de electrónico y, por ende, analizable computacionalmente.

Los términos, objeto de estudio de la terminología, deben estudiarse y recogerse en ejemplos reales de uso; la invención de frases por parte del terminólogo es prácticamente inviable, dado que no trabaja con el lenguaje general que como nativo conoce, sino con el de un ámbito de conocimiento o de actividad, con lo que el lenguaje ahí empleando, en principio, le es ajeno. La recolección de datos (términos, definiciones, contextos) a partir de textos reales se convierte, por tanto, en un modo de proceder fundamental en terminología.

Por otra parte, uno de los factores que puede favorecer un mayor uso de herramientas informáticas en el análisis lingüístico en general, y en el terminológico, en particular, es la existencia de programas flexibles y fáciles de utilizar que, además, estén comercializados, o bien sean de libre distribución. En este sentido, uno de los programas que cumple estas exigencias es *WordSmith Tools* (en adelante, WST), paquete informático desarrollado por M. Scott (1997) de la Universidad de Liverpool y distribuido por Oxford University Press. Cuenta ya con más de ocho años de existencia y su versión actual es la 4.0.

WST está compuesto de: (a) herramientas; y (b) utilidades. Dentro de cada herramienta hay una serie de instrumentos de análisis y de funciones que permiten, entre otras acciones, elaborar listados de palabras monoléxicas, poliléxicas o polilexemáticas[1], de agrupamientos léxicos (*clusters*) —bien de todo el conjunto de textos, o bien de una palabra base—, de palabras claves (*keywords*), por citar unas pocas y algunas de las cuales describimos más detalladamente a continuación. Las herramientas de las que se compone son: *Wordlist*, *KeyWords*, *Concord*. Las utilidades de este programa, por su parte, son una serie extensa de pequeñas aplicaciones dentro del programa que permiten realizar un conjunto variado de acciones para adecuar los ficheros textuales o parte de éstos a nuestros propósitos y necesidades.

Resulta un programa que juzgamos fácil de usar, y pone a disposición del terminólogo una serie de recursos que, bien empleados, son extremadamente útiles en el análisis de varios aspectos del lenguaje especializado objeto de estudio. Entre éstos se encuentran: la identificación de términos simples, de términos compuestos, de colocaciones, la validación de datos, del tema del texto, etc.

En este trabajo perseguimos mostrar nuestra metodología de explotación de córpora de ámbitos profesionales y académicos útil para la fase que en terminología se denomina «de vaciado». Nuestra propuesta metodológica es, a nuestro entender, realizable, en el sentido de que se centra únicamente en las nociones y los medios explotables por un terminólogo no necesariamente experto en informática. Para que tal propuesta sea

[1] Éste es el término empleado habitualmente en la bibliografía sobre lingüística para referirse a una unidad léxica compuesta por dos o más palabras. Otro término compatible es el de *n-grama*, más común en el ámbito del Procesamiento del Lenguaje Natural. Concretamente, se utiliza bi-grama para conjuntos de dos palabras, tri-grama para tres, y así sucesivamente.

eficaz, hemos considerado oportuno integrar en ella tanto los elementos que son necesarios considerar con respecto a las estrategias de recuperación y de análisis textual para el vaciado, como aquéllos referidos a las herramientas empleadas. Ejemplificaremos nuestra metodología de extracción aludiendo al *Corpus de la Piedra Natural* (CPN), constituido para elaborar el *Diccionario de Términos de la Piedra Natural e Industrias Afines* (Alcaraz *et al* 2005).

2. METODOLOGÍA DE EXTRACCIÓN SEMIAUTOMÁTICA

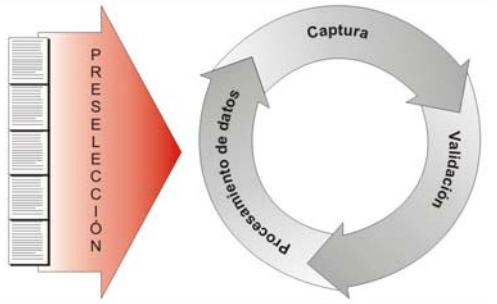


FIGURA 1. Metodología de extracción terminológica

Los córpora especializados que se construyen en cada proyecto terminográfico de nuestro grupo de investigación son analizados por aproximaciones sucesivas, primero en una dirección lingüística (inglés→español) y luego en la otra dirección (español→inglés). Como se observa en la imagen anterior (Figura 1), las cuatro fases de nuestra metodología no son secuenciales, sino que las tres últimas son recursivas, en el sentido de que en muchas ocasiones, y dependiendo del término que se trate de identificar o analizar, debemos volver a la fase anterior para afianzar las informaciones lingüísticas y datos terminológicos que se han obtenido previamente. Asimismo, a medida que se van validando datos en la tercera fase éstos se van procesando y almacenando en la base de datos terminológica creada a tal efecto, por lo que con este método no se pueden disociar de forma nítida la fase tercera y la cuarta, sino que, en realidad, se realizan de forma simultánea. Se expone a continuación una definición más concreta de estas cuatro aproximaciones:

a) La preselección de datos: entendemos esta fase como la primera aproximación al léxico y se basa en el análisis frecuencial. En ella se pretende extraer de forma semiautomática candidatos a término por medio de los listados monoléxicos y poliléxicos que elabora el programa de concordancias empleado, listados que contienen las unidades lingüísticas más frecuentes en los distintos ámbitos profesionales objeto de estudio;

b) La captura de datos: a partir de las listas creadas en la fase anterior se procede a identificar posibles términos a través del examen de los contextos que contienen información útil por medio de los programas de concordancias;

c) La validación de los datos que capturamos en la fase anterior: intentamos encontrar en el corpus más pruebas para decidir si estamos ante una unidad terminológica; se analizan pormenorizadamente los contextos y se identifican en ellos los equivalentes y las definiciones, posibles relaciones de sinonimia, antonimia, se seleccionan fragmentos contextuales, etc.;

d) El procesamiento de los datos: es la última aproximación y en esta fase se van introduciendo los datos en las fichas electrónicas de cada término contenidas en la base de datos.

las cuatro fases de nuestra metodología no son secuenciales, sino que las tres últimas son recursivas, en el sentido de que en muchas ocasiones, y dependiendo del término que se trate de identificar o analizar, debemos volver a la fase anterior para afianzar las informaciones lingüísticas y datos terminológicos que se han obtenido previamente

2.1 Primera aproximación: la preselección de datos mediante la creación de listas por frecuencia

A esta fase la denominamos «de preselección» por dos motivos relacionados con las tareas que aquí se llevan a cabo. El primero es la introducción al programa de un listado de palabras que no queremos que aparezcan ni en los listados por frecuencia ni en los alfabéticos. El segundo motivo tiene que ver con el procesamiento manual que se lleva a cabo para excluir las combinaciones recurrentes sin validez semántica.

La primera operación que se realiza cuando procesamos un corpus es generar un listado de palabras con *WordList*, en la que cada ítem (o *token*) aparece en esta lista ordenado por su frecuencia de aparición y por orden alfabético. Dado que trabajábamos en un proyecto bilingüe estos listados se generaron por cada lengua de trabajo.

Los mecanismos de extracción de términos que utilizan métodos estadísticos, como es el caso de WST, producen muchos datos no válidos o ruido, por lo que se requiere una dedicación humana importante después de haber obtenido los listados de vaciado. En la siguiente tabla (Tabla 1) mostramos las 24 primeras palabras del CPN en inglés y en español. Obsérvese que la primera palabra con contenido semántico y de interés terminológico no aparece hasta la posición 12 en inglés y la 24 en español:

N	Word	Freq.	N	Word	Freq.
1	THE	39641	1	DE	51371
2	OF	23924	2	LA	24889
3	#	16917	3	#	17964
4	AND	15105	4	EN	16762
5	TO	11927	5	Y	16087
6	IN	11383	6	EL	15399
7	A	10691	7	QUE	12015
8	IS	7689	8	SE	10815
9	BE	6010	9	LOS	10534
10	ARE	5227	10	A	10409
11	FOR	5105	11	LAS	10188
12	STONE	4765	12	DEL	7533
13	OR	4466	13	CON	6364
14	WITH	4238	14	POR	5377
15	BY	3494	15	UNA	5188
16	AS	3417	16	PARA	5150
17	ON	3120	17	UN	5042
18	THAT	2812	18	O	4547
19	WHICH	2780	19	ES	3966
20	THIS	2680	20	AL	2983
21	FROM	2565	21	COMO	2903
22	IT	2296	22	SU	2893
23	CAN	1853	23	NO	2581
24	AT	1848	24	PIEDRA	2203

TABLA 1. Listado por frecuencia generado con WST para el CPN en inglés y en español

Para solucionar esta cuestión, WST permite cargar, antes de generar dicho listado, un fichero de texto que contiene una lista de palabras gramaticales, compuesta de artículos definidos e indefinidos, numerales, posesivos, pronombres personales; esto es, una serie que contiene palabras de clase cerrada. Así, uno de nuestros métodos para establecer criterios válidos de identificación de términos consiste en determinar qué palabras se excluyen como términos. Dichas listas[2], que actúan a modo de filtro, son conocidas por el nombre de *stopword lists*; nosotros las denominaremos «filtros

[2] En la siguiente dirección de la Universidad de Neuchatel (Suiza) se pueden descargar filtros léxicos ya elaborados: <http://www.unine.ch/info/clef/>. [Última fecha de consulta: febrero de 2006].

léxicos». Resultan muy útiles para evitar que aparezcan en los listados que producen los programas de concordancias palabras sin contenido específico que, además, salen con una elevada frecuencia en los textos y que, por tanto, generan el indeseado ruido.

2.1.1 Listados monoléxicos

El procedimiento tradicional realizado en el ámbito de la extracción semiautomática de términos es considerar, en primer lugar, los posibles métodos computacionales para elaborar listados monoléxicos. A partir de estos listados ya se puede proceder, posteriormente, con la recuperación de listados compuestos por más de una unidad (bigramas, trigramas, etc.). Los listados por frecuencia que genera la herramienta *WordList* nos muestran posibles términos simples, tal y como se puede observar en la siguiente tabla[3]:

N	Word	Freq.	N	Word	Freq.
1	STONE	5485	1	PIEDRA	2696
2	TEST	2162	2	ROCA	1684
3	MATERIAL	1984	3	ENSAYO	1400
4	ROCK	1896	4	PROBETA	1254
5	USED	1752	5	CANTERA	1123
6	SURFACE	1743	6	MÁRMOL	1105
7	NATURAL	1490	7	TIPO	1093
8	SPECIMEN	1450	8	AGUA	1090
9	USE	1247	9	FORMA	1072
10	MARBLE	1216	10	MM	994

TABLA 2. Listado de las 10 palabras más frecuentes del CPN (inglés y español) con filtro léxico

Estudiar los datos primarios que nos ofrece este tipo de listados proporciona cierta información, principalmente en tres direcciones (Pérez Hernández 2002): en la primera, las listas por frecuencia ofrecen indicaciones sobre si la composición del corpus es adecuada a la temática del ámbito explorado. Si observamos que entre las palabras más frecuentes aparece un número considerable de ítems que no parecen guardar relación con el ámbito especializado objeto de estudio, puede ser que los criterios

[3] La tabla que mostramos contiene las primeras 10 líneas de las listas originales. La original en inglés se compone de 14 567 filas, que traducido a páginas asciende a un total de 232. La original en español cuenta con 22 181 filas, que en páginas es 353.

internos y externos de selección textual[4] necesiten una revisión y ajuste; en la segunda dirección, observar las palabras más frecuentes del corpus nos permite obtener información sobre los campos léxicos y conceptuales más relevantes de nuestro ámbito especializado; en la tercera, a partir de estas palabras más frecuentes podemos observar las líneas de concordancias en las que se hallan insertas, aspecto que permite comprender mejor por qué aparecen con una frecuencia dada y cómo son utilizadas realmente por los especialistas del ámbito productores de los textos que componen un corpus dado.

Los listados generados para cada lengua de trabajo mostraron una elevada frecuencia de aparición de unidades léxicas no especializadas, esto es, de uso corriente, en los textos especializados que componían el CPN. Palabras como *agua*, *medio*, *mayor*, *característica*, *elemento* son unidades del vocabulario general que también forman parte del lenguaje especializado explorado, aspecto que confirma lo difícil que resulta trazar, por lo general, una línea divisoria nítida entre palabras y términos. Es en este tipo de casos que el observar estas unidades funcionando en los contextos donde se hallan se vuelve una cuestión fundamental, dado que es únicamente a través de éstos que podemos llegar a las conclusiones oportunas con respecto a si cierta forma es o no un término.

Los listados generados para cada lengua de trabajo mostraron una elevada frecuencia de aparición de unidades léxicas no especializadas, esto es, de uso corriente, en los textos especializados que componían el CPN

Otro punto interesante de resaltar, y directamente relacionado con lo anterior, tiene que ver con las cuestiones morfológicas de las formas representadas en las tablas anteriores. Si observamos dichas tablas, la mayoría de los ítems son, en principio, candidatos a término por su frecuencia elevada, pero apréciese que la mayor parte procede del lenguaje general. Vemos, por tanto, que las palabras recuperadas como más frecuentes en el CPN, tanto en inglés como en español, no presentan diferencias morfológicas destacables que nos ayuden a distinguirlas de las unidades léxicas generales. En este contexto, podemos afirmar que en nuestro ámbito la identificación y posterior extracción de los términos insertos en los textos resultó ser una tarea no falta de complejidad. En el sector industrial explorado, una gran parte de los términos provienen del lenguaje general, y están, por tanto, desprovistos de patrones terminológicos (por ejemplo, formantes cultos) que faciliten su identificación en el corpus, a excepción de los términos pertenecientes al subdominio exclusivo de la

[4] Los criterios internos se refieren a los rasgos puramente lingüísticos de los textos. En este caso, el investigador examina la recurrencia de patrones lingüísticos (léxicos o sintácticos) en el texto para llevar a cabo una clasificación textual. Los criterios externos son aquéllos de naturaleza esencialmente no lingüística. Así, se derivan del examen de la función comunicativa de los textos, de los interlocutores, de la situación comunicativa y, en definitiva, de los parámetros y las categorías socioculturales (Sinclair 2003, 170).

petrología. A este tipo de unidades Alcaraz (2000) las denomina «términos semitécnicos».

2.1.2 Listados poliléxicos

Con *WordList* es posible también generar listados poliléxicos, a saber: de dos palabras, de tres, de cuatro, hasta un total de ocho. Para ello, en lugar de crear un listado, hemos de seleccionar la opción de <*Make/Add to index*>. Una vez que el índice está elaborado, lo abrimos con *WordList* y entonces 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 (IM). 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* (Tabla 3).

En los listados poliléxicos, tanto si se utilizaba el instrumento de agrupaciones (*clusters*), como si recurríamos al cálculo de la IM, aparecieron multitud de combinaciones de palabras sin ningún interés terminológico, como «de la», «algo como», «algo más», «*by the*», «*is not*», etc. Así, el listado de la IM produjo una lista compuesta por un gran número de páginas. Por tanto, todas estas listas necesitaron una fase de depuración, que se realizó de forma manual; se trataba de depurar estos listados para dejarlos «limpios de ruido». Una vez limpios los listados, la información que proporciona el cálculo de la información mutua se puede visualizar y comprender de un modo más nítido y claro. En este sentido, convenimos que la lista de IM resultó ser, sin lugar a dudas, de suma validez en nuestro trabajo terminológico, pues nos permitió descubrir los verbos, los adjetivos y los adverbios que se combinaban con determinadas unidades seleccionadas. En definitiva, y a nuestro parecer, se revela como el medio más rápido y eficaz para observar los patrones colocacionales o combinatorios de las unidades léxicas. Creemos que en la siguiente selección representada en la Tabla 3 se podrá apreciar esta idea:

	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	Z	MI3	Log L
667	ABRIR	20	CANTERA	513	3	2	4	7,86	9,22	-11,6	568,34
1311	ADMITE	22	PULIDO	206	2	2	4	9,04	14,22	-11,6	171,38
1319	ADMITEN	27	PULIMENTO	52	5	2	5	11,05	32,39	-10,63	8,05
1351	ADQUIEREN	13	BRILLO	61	4	1	5	11,87	43,19	-10,63	33,8
4004	ASERRAR	34	MÁRMOL	1105	1	1	16	7,98	19,36	-5,6	1273,23
5935	CALIZA	249	ORNAMENTAL	137	4	1	7	6,93	8,49	-9,17	32,97
5939	CALIZA	249	GRIS	179	4	1	7	6,55	7,22	-9,17	11,5
5945	CALIZA	249	AZUL	44	1	1	6	8,35	13,55	-9,84	158,3
5947	CALIZA	249	COMPACTA	50	5	1	7	8,39	14,84	-9,17	144,53
5951	CALIZA	249	PACKSTONE	5	1	1	4	10,9	27,5	-11,6	302,94
5952	CALIZA	249	CRISTALINA	50	3	1	3	7,16	6,1	-12,84	144,53

5954	CALIZA	249	BLANCA	34	3	1	7	8,94	18,17	-9,17	184,48
5956	CALIZA	249	BLANDA	22	3	1	3	8,35	9,58	-12,84	223,03
5962	CALIZA	249	FOSILÍFERA	8	2	1	4	10,22	21,67	-11,6	285,02
5963	CALIZA	249	DOLOMITIZADA	5	3	1	5	11,22	34,41	-10,63	302,94
5967	CALIZA	249	DOLOMÍTICA	9	1	1	4	10,05	20,41	-11,6	279,58
5968	CALIZA	249	BRECHOIDE	15	1	1	4	9,32	15,71	-11,6	250,81
5971	CALIZA	249	ARRECIFAL	5	1	1	5	11,22	34,41	-10,63	302,94
5972	CALIZA	249	MARMÓREA	11	7	1	8	10,76	37,06	-8,6	269,33
5973	CALIZA	249	PORTLAND	21	1	1	4	8,83	13,19	-11,6	226,71
5975	CALIZA	249	FRANCESA	10	2	1	5	10,22	24,23	-10,63	274,36
5976	CALIZA	249	IRLANDESA	6	1	1	3	10,22	18,77	-12,84	296,65
5977	CALIZA	249	CRETÁCICA	9	1	1	5	10,37	25,56	-10,63	279,58
5984	CALIZAS	314	ORNAMENTALES	253	11	1	38	8,15	31,72	-1,85	6,58
6000	CALIZAS	314	GRIS	179	2	1	3	4,99	2,11	-12,84	37,44
6007	CALIZAS	314	COMPACTAS	31	5	1	7	8,74	16,9	-9,17	269,75
6012	CALIZAS	314	DOLOMÍTICAS	13	3	1	3	8,77	11,19	-12,84	343,99
6013	CALIZAS	314	LACUSTRES	10	1	1	4	9,57	17,18	-11,6	359,91
6014	CALIZAS	314	BIOCLÁSTICAS	5	3	1	3	10,15	18,3	-12,84	390,75
6015	CALIZAS	314	MARMÓREAS	20	8	1	19	10,81	58,14	-4,85	311,63
6016	CALIZAS	314	RECRISTALIZADAS	8	2	1	5	10,21	24,12	-10,63	371,47
6021	CALIZAS	314	ARENOSAS	9	1	1	6	10,3	27,3	-9,84	365,58
6022	CALIZAS	314	FOSILÍFERAS	8	3	1	7	10,69	33,85	-9,17	371,47
6026	CALIZAS	314	OOLÍTICAS	5	3	1	5	10,89	30,61	-10,63	390,75
6027	CALIZAS	314	CRETÁCICAS	10	4	1	9	10,74	38,93	-8,09	359,91
6028	CALIZAS	314	JURÁSICAS	8	2	1	6	10,47	28,98	-9,84	371,47
6029	CALIZAS	314	CONTINENTALES	7	2	1	4	10,08	20,61	-11,6	377,6
6030	CALIZAS	314	TERCIARIAS	12	2	1	8	10,3	31,53	-8,6	349,13
6031	CALIZAS	314	MICRÍTICAS	8	2	1	4	9,89	19,25	-11,6	371,47
6032	CALIZAS	314	MARMORIZADAS	9	1	1	7	10,52	31,89	-9,17	365,58

TABLA 3. Muestra del CPN (español) con datos de información mutua

En nuestra opinión, los resultados de la tabla anterior son especialmente llamativos. Obsérvese el término *caliza*(*s*). Con este procedimiento se pueden apreciar casi inmediatamente los compuestos de bases características y específicas de un ámbito. Hemos de destacar que el trabajo desarrollado con este listado en las dos lenguas de trabajo resultó ser relativamente sencillo para la detección de combinaciones terminológicas, si bien es cierto que la tarea que más tiempo nos llevó fue la de depuración o eliminación del ruido. Es claro que el material que produce inicialmente está en bruto, pero como punto de partida nos proporcionó mucha y muy buena

información. Podemos afirmar, en este sentido, que los listados que ofrecen la medida estadística de información mutua son de suma validez, por lo que no pueden obviarse y deben, al contrario, ser considerados por el terminólogo como la materia prima de extracción de posibles combinaciones terminológicas. A partir de la anterior tabla pudimos extraer combinaciones como:

- *abrir una cantera; admitir pulido/pulimento; adquirir brillo; aserrar mármol* (V+N);
- *apertura (de) cantera / trincheras / banco / galerías...* (N deverbal+(de)+N); y
- *actividad extractiva / minera / económica / industrial / humana* (N+Adj);
- *caliza ornamental / gris/ azul / compacta / packstone / cristalina/ fosilífera / dolomitizada / brechoide / arrecifal / oolítica....* (N+Adj).

Esta es una fase un tanto mecánica en la que únicamente se preparan los listados eliminando el ruido para poder proceder más adelante con un análisis más exhaustivo de los datos que nos proporcionan las listas. Pasamos, por tanto, a exponer la siguiente fase.

2.2 Segunda aproximación: la captura de datos

Los programas de análisis textual realizan diferentes operaciones que ayudan al terminógrafo a identificar posibles términos junto con sus combinaciones más frecuentes. En la siguiente fase, resultan de suma utilidad, además del listado de información mutua, los de concordancias, en los que seguidamente centraremos nuestra atención.

2.2.1 Listados de concordancias

Si se observan las primeras palabras de una lista monoléxica generada con *WordList*, se pueden identificar los términos centrales de un ámbito especializado. Empleando la terminología de Ahmad y Rogers (2001, 742), se podría decir que estas unidades son los «términos madre» de una especialidad concreta. Como «madres» engendran otros términos a través de los procesos de formación y combinación que sean válidos en la lengua que está siendo objeto de estudio. Así, a partir de los primeros listados generados por frecuencia, y que suponen el punto de partida para la identificación de los términos presentes en el corpus, se «entra» a cada una de las unidades léxicas que aparecen en dicho listado con el programa *Concord*. Los instrumentos que proporciona esta aplicación son: (1) concordancia (*concordance*); (2) lista de colocados (*collocates*); (3) gráfico de distribución de la palabra de búsqueda (*plot*); (4) lista de patrones de colocados (*patterns*); y (5) lista de agrupamientos léxicos (*clusters*).

Con el instrumento *concordance* obtenemos el listado KWIC (*Key Word in Context*) (Figura 2). Este otro método se constituye como un modo sencillo, útil y eficaz para aislar términos poliléxicos a partir de una palabra clave o base.

N	Concordance	Set	Tag	Word No.	File	%
59	solubilizando sus componentes y que comprenden: Agua nebulizada. Como sistema de limpieza es muy			10.560	00421.txt	59
60	probeta seca, se coloca la probeta en un recipiente con agua a una temperatura estándar, se espera un tiempo			1.435	01001.txt	60
61	totalmente metálica. Las almohadillas se llenan con agua a 3,0 MPa de presión, desarrollando un empuje			1.260	00412.txt	20
62	destruyen y tras ese tiempo se lava enérgicamente con agua a presión. Suele ser la primera de otras sucesiv			11.754	00421.txt	66
63	resinas, reptiéndose. Al mes se levanta el papel con agua caliente, si es necesario con el mismo disolvent			15.616	00421.txt	87
64	s métodos de control que se utilizan son: - Riego con agua con o sin estabilizantes químicos. - Pavimentaci			1.247	00415.txt	14
65	raíces. Las baldosas de caliza admiten limpiarse con agua con lejía. Las superficies no deslizantes pueden			4.906	00509.txt	98
66	actuar sobre ambas caras del disco. Se consigue con agua corriente. VI. HERRAMIENTAS UTILIZADAS E			23.394	00801.txt	72
67	(23 ± 5) °C. Finalmente, se lavan cuidadosamente con agua corriente. Las probetas se pesan, después de s			1.308	03901.txt	76
68	quivalen a emplastos, por prepararse amasándolos con agua desionizada o destilada y aplicándolos sobre la			11.070	00421.txt	62
69	ban más burbujas de aire, a continuación, llenarlo con agua desionizada casi hasta el borde y dejar que la			1.329	03701.txt	61
70	continuación, llenar con precaución el picnómetro con agua desionizada hasta el enrase, cerrarlo con su tap			1.361	03701.txt	63
71	cilmente con cualquier fluido. Se pueden mezclar con agua desionizada formando un emplasto para eliminar			10.906	00421.txt	61
72	Vaciar y limpiar el picnómetro, llenarlo únicamente con agua desionizada y pesarlo con una precisión de ± 0,			1.395	03701.txt	64
73	e evitar el uso de abrasivos, el corte se realiza sólo con agua. El coste de los abrasivos puede hacer inviable el			6.199	09301.txt	92
74	y mármol, en rejuntado. RTC-1 Escayola amasada con agua en la proporción de 80 litros de agua por cada 1			1.140	00507.txt	51
75	o y determinar la masa ms de la probeta saturada con agua. En el caso de piedra natural con cavidades visi			1.035	03701.txt	47
76	, en gramos; mz es la masa del picnómetro lleno con agua, en gramos; Vb es el volumen aparente de la pro			420	03701.txt	20
77	nte el trabajo debe estar continuamente refrigerado con agua. En las máquinas corrientes el disco permanece			2.430	05406.txt	45
78	ra eliminar el polvo y material acumulado. . Riego con agua. Es el método más económico, con un grado de			1.271	00415.txt	14
79	empleado, comentemente arena, que, mezclaba con agua, es vertida continuamente en forma de lluvia sobr			1.148	05406.txt	21
80	os, halógenos y álcalis. La limpieza se realizará con agua jabonosa o detergentes no agresivos. Para la c			6.365	00422.txt	72
81	s baldosas de granito y cuarcita podrán limpiarse con agua jabonosa o detergentes no agresivos. Las bal			4.883	00509.txt	97
82	tiempo de actuación y tras levantarlo se pulverizará con agua, lavando así la superficie. Las sales nocivas, los			11.321	00421.txt	64
83	da de espátulas y punzones de madera, quitando con agua los residuos puntuales. Tal actuación es suscepti			10.980	00421.txt	62
84	(polietileno) para evitar la evaporación. Un lavado con agua mediante pulverización y esponja arrastrará cual			11.146	00421.txt	63
85	ez podría vivir sólo tres horas dentro de una pecera con agua normal, pero si el agua es tratada con Bakuhan p			440	00901.txt	94
86	van a estar colocados a la intemperie, en contacto con agua o con la humedad del suelo. 2.1.4. Resistencia			1.359	00405.txt	20
87	e" y las cuñas de madera que las mujeres regaban con agua para que, al dilatarse con la humedad, provocaran			3.720	00801.txt	12
88	gantes con cera y limpiar la piedra frecuentemente con agua para retirar el polvo fino de las cavidades. Desarr			1.011	01201.txt	73
89	vantes que aparecían en la cantera y se remojaban con agua para que, al dilatarse, provocaran el levantamient			6.524	00801.txt	20
90	no requieren ser golpeadas, pues basta regarlas con agua para que, al hincharse, produzcan la separación			1.472	05404.txt	17
91	AL CORTE EN ANGULO FRESADO Y PULIDO CON AGUA. PULIDO INTERNO Foto 15. Diversas operaci			12.103	00413.txt	83

FIGURA 2. Líneas de concordancias KWIC de agua

Si tomamos el ejemplo de la palabra *agua* y la observamos en un listado de concordancias, como el de la figura anterior, podemos extraer términos compuestos y expresiones que se generan a partir de ella. Es el caso de combinaciones léxicas como AGUA *desionizada*, *subterránea/freática*, *potable*, *corriente*, *a presión*, *de corte*, *de escorrentía*, *a dos* AGUAS, *vierte*AGUAS, *absorción de* AGUA *a presión atmosférica*, etc. De este modo, realizando búsquedas a partir de una base o raíz de los primeros términos madre que aparecen en los listados monoléxicos, se pueden identificar patrones estructurales más amplios. La función de búsqueda, además, permite la exclusión de ciertos ítems, por lo que se puede depurar y refinar el procedimiento.

2.2.2 Búsqueda de combinaciones terminológicas

Las búsquedas a partir de las bases sacan a la luz las combinaciones terminológicas en las que determinados términos aparecen. Estas combinaciones unos autores las denominan «colocaciones» y otros «términos compuestos». Al no encontrar un consenso denominativo, ni una distinción tajante sobre si una determinada estructura es un término compuesto, una colocación o una unidad fraseológica especializada, en nuestro trabajo adoptamos una denominación única, esto es, la de «combinación terminológica». La utilizamos como expresión hiperonímica que abarca y alude a todos los tipos de combinaciones que incluyen uno o varios términos y a otras expresiones de interés terminológico.

El procedimiento de búsqueda a partir de bases empleado para objetivos terminológicos no está carente de dificultad. El propósito de una concordancia es, fundamentalmente, identificar colocaciones o proporcionar información sobre «*the company words keep*» (Scott 2003). Sin embargo, desde la perspectiva de la terminología, las líneas de

concordancias tienden a generar tanto falsos términos compuestos, como verdaderos términos. A este respecto, Heid (2001, 791) afirma que «*the relationship and the borderline between collocations [...] and 'multiword terms' is not easy to describ*e». Supera los límites establecidos para este artículo exponer pormenorizadamente las diferencias teóricas entre lo que constituye una colocación, un término compuesto y una unidad fraseológica especializada. Únicamente insistiremos en que el anisomorfismo léxico que encontramos al contrastar dos lenguas, así como los objetivos de nuestra aplicación, nos llevó a adoptar, basándonos en Gläser (1994), un modelo amplio de la combinatoria para abarcar, de este modo, un mayor número de posibilidades combinatorias y fenómenos propios de los lenguajes especializados; para un diccionario terminológico destinado a traductores pensamos que es recomendable recoger, además de los términos simples, las combinaciones terminológicas que puedan suponer en un momento dado un problema de traducción.

Las opciones de análisis se pueden multiplicar con *Concord*, pues podemos observar los términos de nuestra selección en su contexto y, además, es posible indicarle al programa que nos destaque *n* número de palabras a la izquierda de la palabra base y *n* número de palabras a su derecha. Asimismo, y trabajando con los instrumentos *collocates* (Figura 3) y *patterns* (Figura 4), es posible obtener listados de los colocadores o colocativos. Estos listados nos muestran las palabras que aparecen con más frecuencia a la derecha y a la izquierda del término o grupo de palabras que queramos observar.

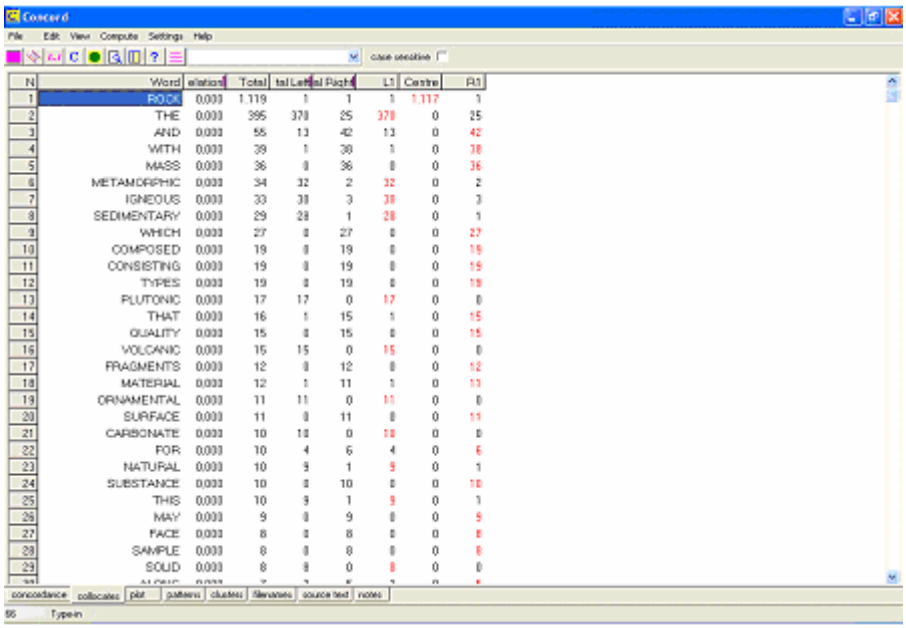
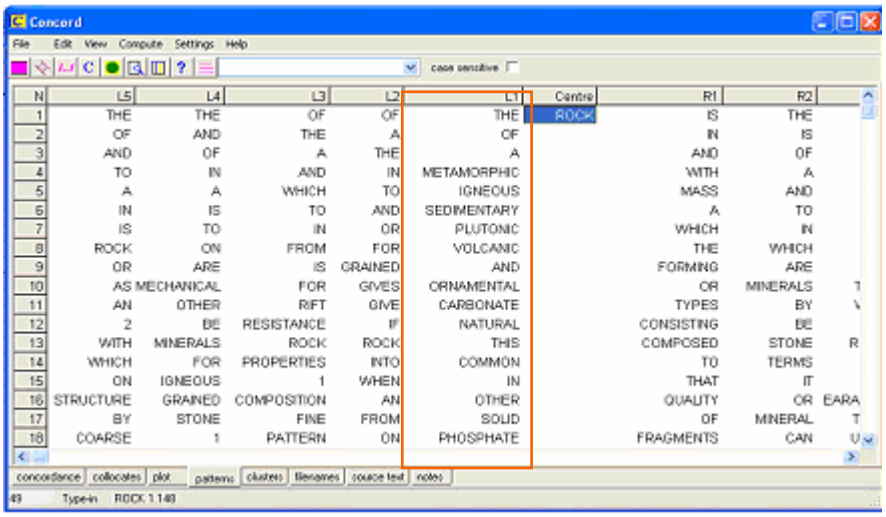


FIGURA 3. Pantalla de *collocates* para la forma *rock*

La figura anterior muestra el listado de palabras que aparecen alrededor de la palabra base ROCK, en posiciones determinadas. La posición de la primera palabra a la derecha de la palabra base es representada por el programa por R1, la segunda por R2, y así sucesivamente. El mismo esquema se aplica a la izquierda: L1 para la primera a la izquierda, L2 para la segunda, etc. En la imagen únicamente aparecen las columnas L1 y R1 porque así hemos configurado esta opción. Por tanto, obtenemos que en la posición izquierda 1 (L1) *rock* aparece frecuentemente (apréciese que destaca en rojo la frecuencia) con *metamorphic*, *igneous*, *sedimentary*, *plutonic*, *carbonate*, *natural*, *solid*, entre otros adjetivos. En la posición derecha 1 (R1) rock funciona como adjetivo y se combina con *mass*, *types*, *quality*, *fragment*, *material*, *surface*, *substance*, *face*, *sample*, etc.



N	L5	L4	L3	L2	L1	Centre	R1	R2
1	THE	THE	OF	OF	THE	ROCK	IS	THE
2	OF	AND	THE	A	OF		IN	IS
3	AND	OF	A	THE	A		AND	OF
4	TO	IN	AND	IN	METAMORPHIC		WITH	A
5	A	A	WHICH	TO	IGNEOUS		MASS	AND
6	IN	IS	TO	AND	SEDIMENTARY		A	TO
7	IS	TO	IN	OR	PLUTONIC		WHICH	IN
8	ROCK	ON	FROM	FOR	VOLCANIC		THE	WHICH
9	OR	ARE	IS	GRAINED	AND		FORMING	ARE
10	AS	MECHANICAL	FOR	GIVES	ORNAMENTAL		OR	MINERALS
11	AN	OTHER	RIFT	GIVE	CARBONATE		TYPES	BY
12	2	BE	RESISTANCE	IF	NATURAL		CONSISTING	BE
13	WITH	MINERALS	ROCK	ROCK	THIS		COMPOSED	STONE
14	WHICH	FOR	PROPERTIES	INTO	COMMON		TO	TERMS
15	ON	IGNEOUS	1	WHEN	IN		THAT	IT
16	STRUCTURE	GRAINED	COMPOSITION	AN	OTHER		QUALITY	OR
17	BY	STONE	FINE	FROM	SOLID		OF	MINERAL
18	COARSE	1	PATTERN	ON	PHOSPHATE		FRAGMENTS	CAN

FIGURA 4. Pantalla de *patterns* para la forma *rock*

Con el instrumento *patterns* (Figura 4) se genera un listado resumen de los colocados, que son agrupados en las posiciones en que aparecen más frecuentemente. Este es otro modo en el que utilizando WST el terminógrafo puede advertir que las unidades léxicas simples son a su vez la base de un término compuesto (*metamorphic* | *sedimentary* | *plutonic* | *volcanic* | *granitic* ROCK), o el colocador de otra base (ROCK *mass* | *quality* | *fragments*). En definitiva, los instrumentos *collocates* o *patterns* de WST son dos modos de llegar a los mismos resultados. La decisión de elegir uno u otro dependerá de con qué listados nos encontremos más cómodos trabajando.

Con el instrumento *clusters* de *Concord* obtenemos un listado de palabras poliléxicas que contienen la palabra base interrogada, aunque el programa también ofrece como resultado los ítems recurrentes en la concordancia, sin limitarse a porciones en los que aparece la palabra de búsqueda (*of the*, *in the*, *test methods*, *determination of*, *name of the*). Sin embargo, algunas de estas agrupaciones léxicas, sin validez desde el punto de

vista terminológico, pueden formar parte de unidades mayores que sí contienen la base (*test methods* aparecía en una agrupación mayor: *natural stone test methods*).

2.3 Tercera aproximación: la validación de datos

Una decisión que se necesita tomar en terminología es qué términos o expresiones se han de recopilar para seguir procesándolos. Los entendidos en terminología y en el discurso especializado coinciden en que las unidades que buscamos aparecen con cierta frecuencia en los textos (*cf.* Ahmad y Rogers 2001, Bowker y Pearson 2002, entre otros). Además de la frecuencia, el corpus construido proporciona a través de los contextos más pruebas que ayudan al terminógrafo a determinar si se halla o no ante un término.

El vocabulario técnico no ofrece mucho problema de identificación y de validación, pues el terminógrafo lo percibe como «exótico», dado que dicho vocabulario no se emplea en el lenguaje general y, por tanto, captará más fácilmente su atención. Es el caso, entre tantos otros, del término *clast* o *clasto*, que, además, aparecía con cierta frecuencia tanto en el CPN en inglés como en el español (Figura 5). Nos pareció obvio que se trataba de una forma muy específica del ámbito de la piedra natural y que, por tanto, convenía procesarla más extensamente para registrar su combinatoria (*sedimento clástico*, *clasto carbonatado*), su serie morfológica (*clástico*), sus compuestos y derivados (*bioclasto*, *bioclástico*, *piroclasto*, *piroclástico*, *intraclasto*, *intraclástico*), las palabras relacionadas (*fragmento*, *grano*, *bloque*, *canto*), las definiciones y sus contextos de uso, entre otros datos.

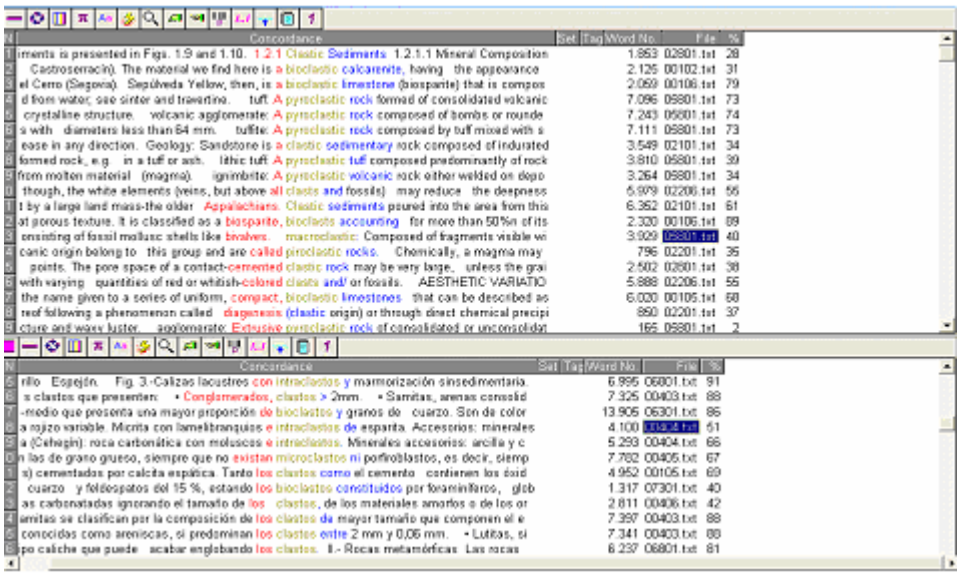


FIGURA 5. Líneas de concordancias KWIC de la búsqueda *clast* en inglés y español

Además del patrón terminológico que presenta la forma *clast→clasto* (se trata de un término procedente del griego *klastos* y que significa «roto, partido»), una detenida observación de los datos nos permite constatar que se trata de una unidad con un significado bien delimitado dentro del ámbito de la piedra natural en general, y del subdominio de la petrología, en particular.

Si bien resulta relativamente fácil identificar determinadas formas como términos, se dan algunos casos un tanto controvertidos. Un buen ejemplo lo constituye la forma *grain→grano*, que es una unidad léxica del lenguaje común que ha adquirido uno o varios nuevos significados dentro del ámbito que tratamos, es decir, las del tipo semitécnico (Alcaraz 2000). Este hecho quedaba constatado observando los contextos. Efectivamente, los ejemplos contextuales nos ayudaron a esclarecer si nos hallábamos ante un término porque, además de mostrar el uso lingüístico de una determinada unidad léxica, éstos nos proporcionaban, en múltiples ocasiones, información conceptual, tanto a través de las definiciones que podían albergar, como a través del uso del término en más de un subdominio de los que constituyen la totalidad del ámbito explorado. De este modo fue como determinamos que *grain* es un término que se usa en petrología, con un significado bien delimitado (*grains are the particles or discrete crystals which comprise a rock or sediment*), en la cantera (*grain in granite is practically the direction in which the stone splits*) y también se utiliza en las herramientas abrasivas (*a grinding disk of grain size F 220*). La unidad léxica *grano* también apareció frecuentemente en el corpus en español y con el mismo uso y significado, a excepción del subdominio de la cantera en donde se no encontró una correspondencia conceptual. Las combinaciones léxicas en las que aparece *grano* como base de una combinación terminológica reafirman el carácter especializado de esta unidad, pues los ejemplos del corpus revelan que dicha combinación vehicula conocimiento específico de distintos subdominios del ámbito explorado. Nos sirve, consecuentemente, para validarlo como término y establecer sus distintos significados.

los ejemplos contextuales nos ayudaron a esclarecer si nos hallábamos ante un término porque, además de mostrar el uso lingüístico de una determinada unidad léxica, éstos nos proporcionaban, en múltiples ocasiones, información conceptual

Son muchos los ejemplos que podemos proporcionar y que ilustran bien esta idea de unidad léxica general que adquiere un valor especializado en contexto (*cabeza→head/crest/top face*; *lecho→bed*; *llaga→head joint*; *luz→span*, etc). Es el caso también de bloque, que puede adoptar hasta cinco significados diferentes dentro del ámbito explorado, como podrá apreciarse en el siguiente fragmento extraído del *Diccionario de Términos de la Piedra Natural e Industrias Afines*:

> **Bloque** *n:* PETRO block; compact aggregate of mineral; S. *agregado ; clasto; fragmento; grano.* [Exp: **bloque** · (PETRO boulder, bowlder; a rock fragment, usually more than 256 mm, and rounded in shape; S. *canto rodado; derrubio*), **bloque** · (PROD rubble stone; natural stone masonry unit, of any shape, with variable dimensions, whose face is rough or worked), **bloque** · (QUAR block; the basic unit of unprocessed stone), **bloque** · (CONST building block) ...

2.3.1 La extracción de conocimiento: recuperación de definiciones, contextos y relaciones conceptuales

Existen determinadas marcas lingüísticas en los contextos que pueden ayudarnos a encontrar bien definiciones, o bien términos relacionados conceptualmente. A este tipo de contextos se les denomina «contextos ricos en conocimiento» (*knowledge-rich context*) (Bowker 1996, Meyer y Mackintosh 1996, Meyer 2001). Se describen como aquellos contextos que identifican al menos una característica conceptual, ya se trate de un atributo o de una relación conceptual concreta (Meyer 2001, 281). Ahmad y Fulford (1992, 7) afirman que las siguientes cinco pruebas diagnósticas de conocimiento (*knowledge probes*) pueden ser utilizadas con un corpus para identificar relaciones de:

1) Sinonimia: la prueba es «X *is* Y»:
2) Hiponimia: aquí las marcas lingüísticas que se proponen para el inglés son «X *is a kind / type / species / shade of* Y»;
3) Meronimia: por ejemplo, «X *is part of* Y»;
4) Causalidad: con «X *causes* Y»; y
5) Material: esto es, «X *is made of* Y».

Podemos establecer una serie compuesta de más pruebas (Tabla 4) al objeto de encontrar definiciones y relaciones conceptuales. En nuestro caso concreto, tuvieron que ser definidas en forma bilingüe, puesto que trabajábamos con córpora en dos idiomas. Los elementos definitorios pueden buscarse en un corpus dado indicando al programa que muestre todos los contextos en donde aparezcan las marcas lingüísticas siguientes, destacadas en negrita, o sus posibles variantes:

Inglés	Español
[X] **are a** [Definition]	[Definición] **se llama** [X]
[X] **(which) is/are called** [Definition]]	[X] **es/son (art.)** [Definición*
[Definition] **is known as (the)** [X]	[Definición] **se conoce como/por el nombre de** [X] [Definición] **conocido/a como** [X]
[Definition] **is named (the)** [X]	[X] **es (art.)** [Definición]
[Definition] **referred to as** [X]	[Definición], **al/ a la que se alude por** [X] [Definición], **referido/a como** [X]
[X] **is defined as** [Definition*	[X] **se define como** [Definición]

TABLA 4. Pruebas diagnósticas para hallar definiciones

Una posible ampliación de las pruebas para detectar las relaciones conceptuales anteriormente citadas y sin ánimo de ser exhaustivos, sino simplemente para dar orientaciones de por dónde se puede trabajar en esta línea con los programas de

concordancias, podría quedar, en los idiomas de nuestro trabajo, de la siguiente manera:

Relación conceptual	Inglés	Español
Sinonimia	also called	también llamado/denominado
	e.g.	p.ej.
	i.e.	esto es, a saber
Hiponimia	a kind of	un tipo de
	a type of	una clase de
	a sort of	una suerte de
	a form of	una forma de
Meronimia	a part of	una parte de
	consists of	consta de/ consiste en
	contains	contiene
Causalidad	causes	causa
	provokes	provoca
	as a result	como resultado
	resulting in	resultando en

TABLA 5. Pruebas diagnósticas para hallar relaciones conceptuales

Con respecto a la determinación de las relaciones conceptuales, en los siguientes ejemplos extraídos del CPN se podrá observar cómo empleando alguna de las pruebas diagnósticas anteriormente referidas (Tablas 4 y 5) pudimos establecer ciertas relaciones. Para obtener resultados de sinonimia pedimos al programa de concordancias que extrajese todos los contextos en donde apareciese la combinación <también llam*>. En los contextos que a continuación presentamos, hemos destacado en rojo la anterior combinación y en azul los términos en los que se establece una relación de sinonimia:

- «Atendiendo a estos criterios se conocen mundialmente los tres grupos denominados genéricamente Granitos, Mármoles y Pizarras, también llamadas Rocas Ornamentales por el valor estético que normalmente lleva aparejado su empleo.»
- «Lo ideal es proceder a la apertura de la cantera con una geometría troncocónica, también llamada en foso, dejando sin extraer una parte del yacimiento [...].»
- «La apertura de un banco cuando existe un talud lateral dentro del hueco de explotación se inicia en uno de sus extremos practicando una trinchera, también llamada triangulada o cajón, que precisará dos planos perpendiculares al frente cortados con hilo diamantado [...].»

Otro caso similar al anterior, pero ahora buscando relaciones de hiponimia, es utilizar la combinación <un tipo de>. Al igual que hicimos con los anteriores contextos, a

continuación resaltamos en rojo la prueba diagnóstica a partir de la que el programa recuperó los contextos en los que aparecía. En azul se destacan los términos entre los que se establece una relación hiponímica:

- «La variedad denominada Blanco Macael es la más importante y característica. Se trata de un tipo de mármol muy blanco, que puede ser elaborado con relativa facilidad en el taller y que se presta a ser esculpido mejor que el de otras canteras.»
- «Un tipo de piedra artificial es el mármol compacto, del que ya hemos hablado en varias ocasiones.»
- «Escoda. Considerada como un tipo de martillo trinchante, con su cabeza terminando en filos paralelos al mango en lugar de bocas. Se utiliza para el acabado final de piezas, dando a la piedra resultante una textura en forma de canales o surcos que tiene como resultado favorecer la escorrentía del agua.»

Vemos que es posible extraer y analizar relaciones conceptuales observando los contextos y sirviéndonos de los programas de concordancias. Con ello, no nos proponemos argumentar que las herramientas informáticas empleadas reconocen directamente dichas relaciones en un corpus, sino que nuestro propósito ha consistido en demostrar que se pueden llevar a cabo análisis rigurosos empleando determinadas marcas lingüísticas relativamente sencillas. El programa de concordancias recuperará los contextos en donde aparezcan dichas marcas y corresponderá al terminógrafo evaluar, en última instancia, si los datos que proporcionan los ejemplos contextuales arrojan algo de luz sobre las posibles relaciones conceptuales entre dos términos dados.

2.3.2 La búsqueda de equivalentes

Otra prueba que nos ayuda a determinar si nos hallamos frente a un término es el encontrar su equivalente en el corpus creado para el otro idioma. Para establecer las correspondencias entre un término de una lengua A con otro de la lengua B partimos de la base de que nuestros córpora son temáticamente comparables[5], y que, además, contamos con textos paralelos o traducidos (A→B; B→A) incluidos en los córpora.

[5] Los córpora bilingües o multilingües se pueden categorizar en: (a) córpora comparables; y (b) paralelos (también denominados «de traducción»). Un corpus comparable es aquél que está compuesto de dos o más córpora en diferentes lenguas (p. ej., inglés y español) o variedades diferentes de una lengua (p. ej., inglés británico, canadiense, estadounidense, etc.). Los textos que componen cada uno de los córpora se seleccionan porque presentan determinados rasgos o características comunes. Las características que suelen compartir son: el tema, el tipo de texto, el periodo de tiempo en que se redactaron los textos, la función comunicativa, el grado de especialización, etc. Un ejemplo que ilustrase bien esta idea sería el caso de un corpus compuesto por artículos científicos en dos o más lenguas (tipo textual compartido) escritos en los últimos diez años (período de tiempo compartido) que versasen sobre la restauración de la

A modo de ilustración, se nos dio el caso de que en los textos incorporados al programa de concordancias que compartían la característica de pertenecer al subcampo de «seguridad» aparecieron con una frecuencia alta unidades léxicas en inglés como *machine*, *sound*, *noise*, *safety*, *hazards*, entre otras. En el subcorpus comparable en español observamos también estas mismas unidades (*máquina*, *ruido*, *seguridad*, *riesgo*, *peligro*), dentro, además, de un intervalo similar de frecuencia. De este modo, frecuentemente podemos encontrar el término equivalente del corpus A en el corpus B.

Con respecto a los subcorpora paralelos, y siempre que consideremos que los textos contenidos en este tipo de corpus nos proporcionan la suficiente fiabilidad, el terminógrafo puede plantearse trabajar exclusivamente con ellos y extraer de ahí todos sus términos y combinaciones terminológicas equivalentes. Para ello, es posible emplear las mismas herramientas e instrumentos que proporciona WST, pero ahora trabajando de forma simultánea en paralelo (véase Figura 5). El uso de córpora paralelos y de herramientas para su explotación puede ayudar a solventar los problemas de búsqueda de equivalentes que plantea el entorno de trabajo terminológico bilingüe (*cf.* Gómez y Vargas 2003).

Con respecto a los subcorpora paralelos (…) el terminógrafo puede plantearse trabajar exclusivamente con ellos y extraer de ahí todos sus términos y combinaciones terminológicas equivalentes

2.4 Cuarta aproximación: el procesamiento de los datos

Por último, en esta fase, que recordemos no separábamos tajantemente de la anterior, tratábamos de procesar y gestionar todos los datos obtenidos y ya validados en la etapa previamente descrita (términos, combinaciones, contextos, definiciones y relaciones conceptuales) en una base de datos terminológica. Se trataba, en definitiva, de ir alimentando la ficha electrónica para ir dando forma y confeccionar así los artículos terminográficos definitivos, como mostramos en forma gráfica a continuación:

piedra natural (tema compartido) y cuya función también es compartida (función informativa). Es claro que cuantas más características compartan los córpora mayor será el grado de correspondencia entre ellos. El corpus paralelo, por su parte, consiste en un conjunto de textos redactados en una lengua (la original) junto con sus traducciones a otras lenguas.

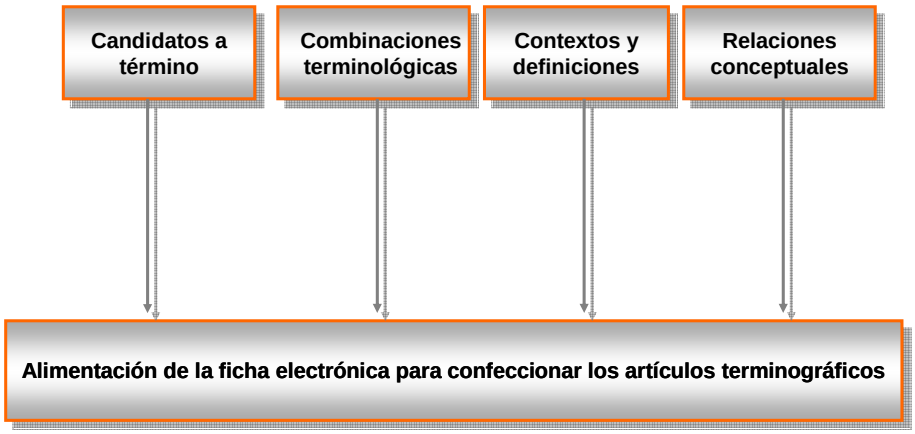


FIGURA 6. Procesamiento de los datos (extraída de Vargas 2005)

Esta es una fase un tanto compleja por la multitud de tareas que se realizaban al mismo tiempo y por la necesidad de tener que interactuar con otras aplicaciones informáticas y consultar otros recursos (enciclopédicos, terminológicos y lexicográficos) disponibles, principalmente, en Internet. De este modo, además del programa de concordancias que utilizamos en la fase anterior, empleamos, de forma combinada, la base de datos terminológica, recurrimos a bancos de datos y diccionarios en línea y accedíamos a Internet para obtener otras informaciones (equivalentes, definiciones, contextos más ilustrativos) que el corpus no nos proporcionaba. Un corpus no es un recurso infalible, tiene sus limitaciones, y en ocasiones determinados conceptos no quedan lo suficientemente claros observando todos los contextos que recupera el programa de análisis textual que se emplee. En otros casos, la carencia de nuestro corpus se reflejaba en la imposibilidad de encontrar en él los equivalentes de unos términos o combinaciones terminológicas dadas, por lo que tuvimos que consultar otras herramientas de naturaleza terminológica o lexicográfica que componían, en definitiva, de forma adicional nuestra estación de trabajo.

A continuación, en el siguiente gráfico ofrecemos una síntesis de lo que constituyó, en el marco del proyecto terminológico emprendido para elaborar el diccionario mencionado en la introducción, todo el proceso de extracción terminológica bilingüe basada en corpus:

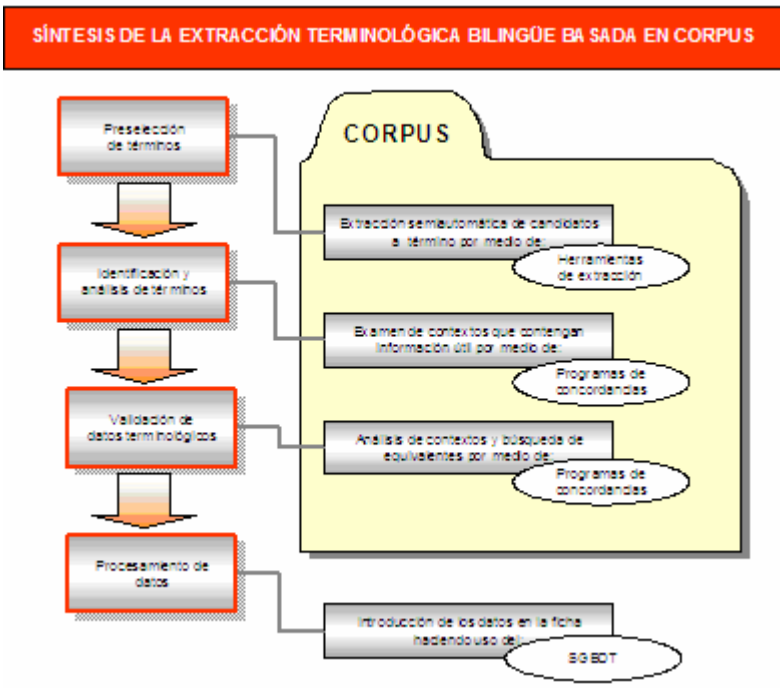


FIGURA 7. Extracción terminológica bilingüe basada en corpus (extraída de Vargas 2005)

3. Conclusión

Este artículo ha sido consagrado a cuestiones sobre la detección y extracción de terminología. Hemos tratado diferentes aspectos directamente relacionados con esa cuestión y, para ejemplificar los distintos procesos, hemos recurrido al CPN. Su meta ha sido proponer una metodología para la explotación sistemática y semiautomática de un corpus especializado. Hemos configurado dicha propuesta sobre cuatro aproximaciones, que son: la preselección, la captura, la validación y el procesamiento de los datos. Asimismo, hemos ofrecido una forma metódica de utilizar unas determinadas herramientas informáticas, junto con los instrumentos que éstas incorporan y hemos abogado por la utilización de una serie de estrategias, que se pueden resumir en: (1) la observación de los índices de frecuencia; (2) la consideración de la forma simple o combinada de los términos que proporcionan los listados; y (3) el hallazgo de determinadas marcas en los contextos para detectar definiciones o términos relacionados semánticamente.

En conclusión, y como se puede apreciar en la figura anterior, un corpus electrónico y las herramientas informáticas que se emplean durante su explotación son, en conjunto, un recurso de primer orden y de suma utilidad que, de hecho, en nuestro proyecto terminológico intervino en la mayor parte de las aproximaciones metodológicas que hemos propuesto. Así, creemos que ha quedado suficientemente constatado que este recurso textual informatizado se convierte en un útil imprescindible en la preselección de los términos, en su identificación y análisis, así como en la validación de datos terminológicos. Es por esta razón que nosotros consideramos que todo proyecto terminológico debe prever una etapa diferenciada y bien planificada en la que se explote el corpus a partir del que se van a extraer todas las informaciones de interés terminológico. ■

5. Referencias bibliográficas

Ahmad, K. y M. Rogers. "Corpus Linguistics and Terminology Extraction". En *Handbook of Terminology Management*, vol. 2, eds. S. E. Wright y G. Budin, 725-760. Amsterdam/Filadelfia: John Benjamins, 2001.

Ahmad, K. y H. Fulford. *Knowledge Processing 5. Discourse Structures and their Use in Elaborating Terminology*. Department of Mathematical and Computing Sciences: University of Surrey, CS-92-08, 1992.

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

Alcaraz, E., Hughes, B., Mateo, J., Vargas, Ch. y A. Gómez. *Diccionario de Términos de la Piedra Natural e Industrias Afines (Inglés-Español, Spanish-English)*. Barcelona: Editorial Ariel, 2005.

Bowker, L. "Towards a Corpus-Based Approach to Terminography". *Terminology* 3, n.º 1 (1996): 27-52.

Bowker, L. y J. Pearson. *Working with Specialized Language. A practical guide to using corpora*. Londres/Nueva York: Routledge, 2002.

Gómez, A. y Ch. Vargas. "Una herramienta de traducción asistida: la aplicación Multiconcord en la extracción de terminología bilingüe". En *Terminología y traducción: un bosquejo de su evolución*, ed. N. Gallardo San Salvador, 227-241. Granada: Atrio, 2003.

Gläser, R. "Relations between Phraseology and Terminology with Special Reference to English". *Alfa: Actes de langue française et de linguistique*, vol.7/8 (1994/5): 41-60.

Heid, U. "Collocations in Sublanguage Texts: Extraction from Corpora". En *Handbook of Terminology Management*, vol. 2, eds. S. E. Wright y G.Budin, 788-803. Amsterdam/Filadelfia: John Benjamins, 2001.

Meyer, I. "Extracting knowledge-rich contexts for terminography. A conceptual and methodological framework". En *Recent Advances in Computational Terminology*, eds D. Bourigault, Ch. Jacquemin y M.C. L'Homme, 280-302. Amsterdam/Filadelfia: John Benjamins, 2001.

Meyer, I. y K. Mackintosh. "The Corpus from a Terminographer's Viewpoint". *International Journal of Corpus Linguistics* 1, n.º 2 (1996): 257-285.

Pérez Hernández, Ch. 2002. "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento." [En línea]. Tesis Doctoral. Universidad de Málaga. Consultado el 3 de febrero de 2006. Disponible en <http://elies.rediris.es/elies18/>.

Sinclair, J. "Corpora for lexicography". En *A Practical Guide to Lexicography*, ed. P. Sterkenburg, 167-178. Amsterdam/Filadelfia: John Benjamins, 2003.

Scott, M. *WordSmith Tools version 4.0*. Oxford: Oxford University Press, 2003.

Vargas Sierra, Ch. Aproximación terminográfica al lenguaje de la piedra natural. Propuesta de sistematización para la elaboración de un diccionario traductológico. Tesis Doctoral. Alicante: Universidad de Alicante, 2005.