



CENTRO DE LINGÜÍSTICA APLICADA
MINISTERIO DE CIENCIA, TECNOLOGÍA Y MEDIO AMBIENTE
SANTIAGO DE CUBA

ACTAS-I

X SIMPOSIO INTERNACIONAL **COMUNICACIÓN SOCIAL**

20 AÑOS
DE COMUNICACION
CIENTIFICA

SANTIAGO DE CUBA 22-26 ENERO 2007

Edición: *Leonel Ruiz Miyares, Alex Muñoz Alvarado y Celia Álvarez Moreno*

Cubierta: *Emilio Pérez Yero*

© Sobre la presente edición:
ACTAS – I, X Simposio Internacional de Comunicación Social
Santiago de Cuba, 22-26 de enero del 2007
Centro de Lingüística Aplicada, 2007

ISBN: 959-7174-08-1

CENTRO DE LINGÜÍSTICA APLICADA
Delegación Territorial del CITMA
Apartado Postal 4067. Vista Alegre.
Santiago de Cuba. Cuba C.P. 90400
Correo electrónico: leonel@lingapli.ciges.inf.cu

ISBN 959-7174-08-1



9 789597 174080

COAUSPICIADORES

- Universidad de Oriente, Santiago de Cuba, Cuba
- Universidad Pedagógica "Frank País García", Santiago de Cuba, Cuba
- Sindicato Nacional de Trabajadores de las Ciencias
- Dirección Provincial de Cultura, Santiago de Cuba, Cuba
- Centro Cultural Africano "Fernando Ortiz", Santiago de Cuba, Cuba
- Centro de Convenciones "Heredia", Santiago de Cuba, Cuba
- Universidad Central de Las Villas, Villa Clara, Cuba
- Universidad de Twente, Enschede, Países Bajos
- Digital Enterprise Research Institute, Leopold-Franzens Universität, Innsbruck, Austria
- Universidad del País Vasco, España
- Grupo de Investigación de Lexicología / Terminología, Universidad Libre de Amsterdam, Países Bajos
- Universidad de Málaga, Málaga, España
- Universidad Rovira i Virgili, Tarragona, España
- Universidad de Alicante, España
- Universidad de Wolverhampton, Reino Unido
- Universidad de Sheffield, Reino Unido
- Universidad de Sunderland, Reino Unido

COMITÉ ORGANIZADOR Y CIENTÍFICO

Pedro Aníbal Beatón Soler	Ministerio de Ciencia, Tecnología y Medio Ambiente
Eloína Miyares Bermúdez	Centro de Lingüística Aplicada
Vitelio Ruiz Hernández	Centro de Lingüística Aplicada
Leonel Ruiz Miyares	Centro de Lingüística Aplicada
Zaida Valdés Estrada	Universidad de Oriente
Ena Elsa Velázquez Cobiella	Universidad Pedagógica "Frank País García"
Anton Nijholt	Universidad de Twente
Nancy Cristina Álamo Suárez	Centro de Lingüística Aplicada
Celia Pérez Marqués	Centro de Lingüística Aplicada
Alex Muñoz Alvarado	Centro de Lingüística Aplicada
Mercedes Cathcart Roca	Universidad de Oriente
Ercilia Estrada Estrada	Universidad de Oriente
Martha Cordiés Jackson	Centro Cultural Africano "Fernando Ortiz"
Miladys Diodene Adame	Ministerio de Educación
Dieter Fensel	Leopold-Franzens Universität
Iñaki Alegría Loinaz	Universidad del País Vasco
Arantza Díaz de Ilarraza	Universidad del País Vasco
Xabier Artola Zubillaga	Universidad del País Vasco
Xabier Arregi Iparragirre	Universidad del País Vasco
Gloria Corpas Pastor	Universidad de Málaga
Daniela Ratti	Instituto de Lingüística Computacional
Lucía Marconi	Instituto de Lingüística Computacional
Paola Cutugno	Instituto de Lingüística Computacional
Ruslan Mitkov	Universidad de Wolverhampton
Federico Albano Leoni	Universidad de Roma 1
Massimo Pettorino	Università L'Orientale
Celia Álvarez Moreno	Centro de Lingüística Aplicada
Mileidis Quintana Polanco	Centro de Lingüística Aplicada
Matilde Moltó Martorell	Centro de Convenciones "Heredia"

- María Rosario Bautista Zambrana
Natural Language Generation and Translation Technologies /377
- Rosa M. Rodríguez Miniet
Algunos fraseologismos cubanos y venezolanos: una sola intención comunicativa /382
- Sandra Sovilj-Nikic y otros
Analysis of different factors influencing vowel duration in the Serbian language /385
- Sophie Kern y Frederique Gayraud
Influence du sexe sur l'acquisition des premiers mots /390
- Stefan Barme
Los galicismos léxicos y sintácticos de la variedad cubana del español /394
- Steven Byrd
Calunga: un dialecto afro-brasileño de Minas Gerais – un breve análisis gramatical /398
- Susana Cisneros Garbey y Rosa Rodríguez Miniet
Diccionario Básico Escolar Cubano. Edición electrónica: una vía para la comprensión textual /404
- Vitelio Manuel Ruiz Miyares
La realidad de los derechos lingüísticos en el método de alfabetización "Yo, sí puedo" /407
- Willy Martin
The Lexicon is a (kind of) Frame /410
- Wim Vandenbussche
Shared Standardization Factors in the History of 16 Germanic Languages /419
- Yaquelin Fonseca Arranz
El conocimiento léxico y su trascendencia sociocultural /424
- Yishai Tobin
A Semiotic View of Signed versus Spoken Language /428
- Yolanda G. López Franco
Los nombres de pila de quienes nacieron entre 1960 y 1975 en Tlalnepantla de Baz, Estado de México. Algunos usos sociolingüísticos /433
- Zoila Guerra Mastrapa y otros
La enseñanza del idioma español: una necesidad imperiosa /438
- Comisión: Lingüística Computacional** /443
- Amparo Alcina y Victoria Soler
Procedimientos y programas informáticos para la extracción automática de términos: estudio preliminar en un corpus del ámbito de la cerámica /445
- Mariët Theune y otros
Questions, Pictures, Answers: Introducing Pictures in Question-Answering Systems /450
- T. Amghar y otros
Using genetic algorithms to compose documents: A way to focus attention on relevant information /464
- Boris Van Schooten y Rieks Op Den Akker
Multimodal follow-up questions to multimodal answers in a QA system /469
- Elina Lagoudaki
Challenges and Possibilities for Extracting Parallel Corpora from the Web – The Translator's Dream Scenario /474
- Gloria Corpas Pastor y Miriam Seghiri Domínguez
Surfing the Net: an R&D Project on Tourism Contracts /480
- Giovanna Morgavi y otros
Instruments for evaluating communication processes /485

AMPARO ALCINA
VICTORIA SOLER
Universitat Jaume I de Castelló
España
alcina@trad.uji.es

Procedimientos y programas informáticos para la extracción automática de términos: estudio preliminar en un corpus del ámbito de la cerámica

Resumen

En el presente artículo se exponen los resultados obtenidos en el marco del proyecto TXTCerám en relación con la extracción automática de términos. A partir de un corpus electrónico de textos de especialidad del ámbito de la cerámica se ha puesto a prueba la eficacia de algunas herramientas informáticas para diseñar un sistema integral de terminología asistida que sea útil para la elaboración de terminologías y para la consulta y uso por mediadores lingüísticos, especialmente traductores. Concretamente, se han utilizado la herramienta de extracción terminológica ExtraTerm de Trados y la herramienta de análisis textual WordSmith. Los resultados de esta extracción automática se comparan con los obtenidos en una extracción manual en un proyecto anterior.

1. Introducción

La globalización de los mercados y las relaciones internacionales y la diversificación continua de los campos del saber y especialización de las disciplinas tiene como consecuencia el crecimiento notable de los términos en los distintos campos del saber y también su continua evolución de su significado. En estas circunstancias, es necesario desarrollar procedimientos y técnicas que permitan tanto la extracción de los nuevos términos que sea rápida y que a su vez mantenga criterios de calidad. Para conseguir este objetivo se hace necesario contar con métodos y recursos tecnológicos avanzados.

En el marco del proyecto TXTCerám¹ hemos observado ventajas e inconvenientes y evaluado la eficacia de algunas herramientas informáticas de extracción automática de términos y su utilidad para la consulta y uso por mediadores lingüísticos, especialmente traductores. En nuestro caso, los resultados de la extracción de términos e información conceptual contribuirán a la creación de la base de conocimiento OntoCerám, en el marco del proyecto ONTODIC² en el que además, se pretende diseñar una metodología de trabajo que sea útil al terminólogo y también al traductor.

En este artículo describiremos los pasos que hemos llevado a cabo para comprobar las posibilidades de la herramienta WordSmith (en concreto WordList) para extraer terminología de nuestro corpus. A continuación describiremos brevemente el corpus que hemos elaborado en el marco del proyecto y que hemos utilizado para probar las herramientas de extracción. En los apartados siguientes describiremos cómo hemos utilizado la herramienta WordSmith y los recursos que hemos creado para poder utilizarla. Finalmente, expondremos las ventajas e inconvenientes que tiene para un traductor.

2. El corpus electrónico de textos del ámbito de la cerámica industrial TXTCerám

Como resultado de proyectos anteriores,³ contamos con una selección de 25 obras especializadas, entre libros, manuales y revistas monográficas, y también algunos folletos de carácter divulgativo o comercial. Se desecharon aquellas que no estaban escritas en español, y aquellas cuya temática era excesivamente específica para nuestros propósitos. Finalmente, con ayuda de los especialistas, se eliminaron también aquellas referencias cuyo contenido no respondía específicamente al campo de la cerámica tradicional. Sin embargo, en anteriores proyectos se había trabajado con los documentos impresos de estas obras.

El proyecto TXTCerám contemplaba necesariamente la creación del corpus electrónico de textos de especialidad del ámbito de la cerámica TXTCerám. La fase-inicial del proyecto consistió en la digitalización de las obras impresas y su organización en un corpus textual de manera que pudiera ser procesado automáticamente mediante el ordenador. Para ello, contamos con la colaboración de estudiantes de las asignaturas de Informática aplicada a la Traducción y Terminología, de la licenciatura en Traducción e Interpretación, de estudiantes de doctorado y de becarios del proyecto. Estos trabajos de digitalización se enmarcaron en un proyecto de innovación educativa que se proponía mejorar las habilidades informáticas de estos estudiantes y aumentar el tiempo productivo de trabajo frente al ordenador (proyecto CREC). Los estudiantes aprendieron a usar el escáner, el programa de reconocimiento óptico de caracteres OmniPage y también tuvieron que familiarizarse con las herramientas de corrección ortográfica y gramatical de Microsoft Word (Soler Puertes et al. 2005). Por otra parte, sus trabajos se incorporaban al corpus TXTCerám.

¹ TXTCerám: Extracción semiautomática y análisis conceptual formal de términos de la cerámica a partir de un corpus electrónico. Su eficacia y utilidad en la mediación lingüística es un proyecto financiado por la Generalitat Valenciana (GV05/260).

² ONTODIC: Metodología y tecnologías para la elaboración de diccionarios onomasiológicos basados en ontologías. Recursos terminológicos para la e-traducción es un proyecto financiado por el Ministerio de Educación y Ciencia de España (TS12006-01911).

³ Estudio y descripción de la lengua de especialidad de una rama profesional de la cerámica y elaboración de un diccionario terminológico multilingüe, financiados por la Fundació Caja de Castellón-Bancaja (P1A98-12) y Generalitat Valenciana (GV00-143-9), para el desarrollo de la fase I (1998-2000) y fase II (2001-02), respectivamente.

En esta fase del proyecto se consiguió recopilar el corpus TXTCeram, en el que constan 28 referencias bibliográficas del dominio de la cerámica industrial en español. El corpus contiene casi 2'5 millones de palabras (exactamente 2.340.161 de palabras), distribuidas en 114 ficheros. Los ficheros están disponibles tanto en formato de texto plano (formato txt) como en formato de Microsoft Word (formato con la extensión doc).

3. Extracción automática con WordSmith

WordSmith (Scott) es un paquete informático de análisis textual y lexicográfico creado por Mike Scott que contiene distintas herramientas (WordList, Concord, KeyWords) y diversas funciones en comparación con otros programas del mismo tipo más antiguos (Alcina y Pruñonosa Tomás 1992). Se trata de una herramienta más orientada a la investigación lexicográfica, por lo que es necesario adaptarla a los fines que cada investigador pretenda alcanzar. WordList es una herramienta que permite elaborar listados de las palabras de un corpus ordenados por su frecuencia de aparición en el corpus o alfabéticamente. También proporciona datos estadísticos sobre algunas características de esas palabras (su longitud, su distribución a lo largo del corpus, etc.). Este puede ser un buen punto de partida para localizar fácilmente los términos de un ámbito de especialidad. Algunos autores han señalado su utilidad en terminología (Faber y Jiménez 2002; López Rodríguez 2001) y en la traducción profesional (Sánchez-Gijón 2005).

3.1. Preparación de la extracción con WordSmith

Los listados de palabras con WordList no son más que la ordenación alfabética y recuento de todas las palabras de los textos, sin discriminar las palabras léxicas de las palabras gramaticales. Además, los recuentos de palabras son recuentos de las distintas formas en las que puede aparecer una misma palabra. Por ejemplo, la palabra *aldosa* puede aparecer en un texto como: *aldosa* y como *aldosas*. Pues bien, para el programa *aldosa* y *aldosas* son dos palabras distintas. Lo mismo ocurre con los verbos. Las distintas formas del verbo *esaltar*: *esaltada*, *esaltan*, *esalta*, aparecen como palabras distintas. Para sacar un mejor provecho del programa hemos contado con algunas funciones que ofrece, que son la stoplist y la lematización, que vemos a continuación.

Eliminación de las palabras gramaticales

Para eliminar las palabras gramaticales del listado que ofrece automáticamente el programa, se elaboró un listado de palabras gramaticales del español en el que constan: determinantes (artículos, demostrativos, posesivos, indefinidos, interrogativos), preposiciones, pronombres (personales, demostrativos, numerales, indefinidos, interrogativos/exclamativos, relativos), adverbios (de lugar, de tiempo, de manera, de afirmación, de negación, etc) y conjunciones.

Para ello, se escogió un libro de Lengua española de nivel bachillerato y se tomaron las clasificaciones que allí aparecían. Todas estas palabras se dispusieron en el formato exigido por WordSmith para crear este tipo de ficheros. En concreto, las palabras debían aparecer en mayúsculas y separadas por comas en un fichero de texto plano (formato txt). En total, el fichero de stoplist carga 306 palabras.

Agrupación de palabras flexionadas correspondientes a una misma forma canónica

WordList permite que las palabras flexionadas correspondientes a una misma forma canónica aparezcan bajo una misma forma. Para ello, es necesario crear un fichero que contenga las agrupaciones de esas palabras. Por ejemplo, para indicar que todas las formas flexionadas del verbo traducir deben unirse bajo una misma forma, debemos escribir todas las formas flexionadas en un mismo párrafo separadas por comas. La primera palabra (normalmente el infinitivo del verbo) debe aparecer al principio y a continuación el símbolo guión '-' y mayor que '>' como en el ejemplo:

```
traducir ->
traduciendo, traducido, traducida, traducidos, traducidas, traduzcas, traduzca,
traduzcamos, traducid, traduzcan, traduzco, traduzcas, traduzca, traducimos, tra
ducís, traducen, traducía, traducías, traducía, traducíamos, traducíais, traduci
an, traduje, tradujiste, tradujo, tradujimos, tradujisteis, tradujeron, traducir
é, traducirás, traduciremos, traduciréis, traducirán, traduciría, traducirías, t
raduciría, traduciríamos, traduciríais, traducirían, traduzca, traduzcas, tradu
zca, traduzcamos, traduzcáis, traduzcan, tradujera, tradujeras, tradujera, tradu
jéramos, tradujerais, tradujeran, tradujese, tradujeses, tradujese, tradujésemo
s, tradujeseis, tradujesen
```

Para crear un listado completo se utilizó un libro de flexión verbal del español, en el que aparecían, por una parte, se flexionaban todos los modelos de verbo flexionados, y por otra parte, un listado de todos los verbos del español con indicación del modelo según el cual debían flexionarse. El fichero de lematización creado contiene 5.434 formas.

3.2. Elaboración de listados de palabras

Una vez tenemos el corpus TXTCeram, la stoplist y el fichero de lemas, configuramos WordSmith para elaborar los listados de palabras. En la figura 1 vemos una pantalla de WordList en la que podemos ver, en la primera columna, las 33 primeras palabras con mayor frecuencia de aparición en el corpus. En la segunda columna se

muestra el número de apariciones de la palabra junto con la de sus formas flexionadas. A continuación, en la tercera columna vemos la frecuencia relativa de esa palabra respecto al resto del corpus. Por último, la última columna muestra las formas flexionadas que se han agrupado junto a la forma base principal que aparece en la primera columna, con indicación del número de ocurrencias de cada una de esas palabras.

Frecuencia	Frecuencia relativa	Formas flexionadas
32 314	0.32	ser(915) sido(1079) soy(4) eres(1) es(16615) acemos(1) sos(4) van(5991)...
10 970	0.45	post(107) po(43) pu(9) pu(9) pu(9) pu(9) pu(9) pu(9) pu(9) pu(9) pu(9) pu(9)...
9 554	0.30	ndada(265) ndada(144) ndada(295) ndada(119) ndada(5) ndada(252) nd...
6 156	0.30	formado(277) formado(26) formado(191) formado(38) formado(124) forma(5)
5 013	0.23	construido(4) construido(3) construido(2) construido(1) construido(1)...
5 514	0.23	...
4 259	0.20	esmal(1) esmal(1) esmal(1) esmal(1) esmal(1) esmal(1) esmal(1) esmal(1)...
4 252	0.20	temido(235) temido(105) temido(4) temido(1) temido(1) temido(1) temido(1)...
4 754	0.20	le(1) le(1) le(1) le(1) le(1) le(1) le(1) le(1) le(1) le(1) le(1) le(1) le(1)...
4 750	0.19	figur(1) figur(1) figur(1) figur(1) figur(1) figur(1) figur(1) figur(1) figur(1)...
4 711	0.19	partido(1) partido(1) partido(1) partido(1) partido(1) partido(1) partido(1)...
4 276	0.17	de(1) de(1) de(1) de(1) de(1) de(1) de(1) de(1) de(1) de(1) de(1) de(1) de(1)...
4 156	0.17	...
4 276	0.17	presentado(57) presentado(41) presentado(24) presentado(20) presentado(15)
3 55	0.15	...
3 459	0.14	...
3 364	0.14	...
3 275	0.13	caso(1) caso(1) caso(1) caso(1) caso(1) caso(1) caso(1) caso(1) caso(1)...
3 224	0.13	secund(1) secund(1) secund(1) secund(1) secund(1) secund(1) secund(1)...
3 222	0.13	...
3 255	0.13	conten(1) conten(1) conten(1) conten(1) conten(1) conten(1) conten(1)...
3 146	0.13	conten(1) conten(1) conten(1) conten(1) conten(1) conten(1) conten(1)...
2 276	0.12	...
2 266	0.12	...
2 344	0.12	...
2 765	0.11	capas(1) capas(1) capas(1) capas(1) capas(1) capas(1) capas(1)...
2 765	0.11	...
2 775	0.11	...
2 765	0.11	...
2 765	0.11	mostrado(4) mostrado(4) mostrado(2) mostrado(2) mostrado(1) mostrado(1)...

Figura 1. Pantalla de WordList con las palabras de mayor frecuencia en TXTCerám

3.3. Análisis de los resultados

A partir del listado de palabras ordenado de mayor a menor frecuencia, hemos estudiado las 100 primeras palabras para comprobar: 1) si constituyen palabras del léxico (no palabras gramaticales) y 2) si constituyen términos de la cerámica industrial. De este análisis es necesario llamar la atención sobre los siguientes aspectos:

- Palabras ambiguas: aparecen algunas palabras en las que se lematizan formas que corresponden a un verbo y también a un sustantivo. Es el caso de *prima* que integra las formas: *prime*, *primo*, *primas*, *primaba*, *primaron*, *primasen*. Podemos utilizar la herramienta Concord para resolver estas dudas. Véase el caso de *prima*: Al analizar todos los casos,
 - *prima* aparece siempre complementando a materia y formando así el término materia.prima (346 apariciones)
 - *primo* aparece en dos ocasiones, como nombre propio (Primo) y como adjetivo relacional
 - *primaba*, *primaron* y *primasen*, del verbo *primar*, aparecen una vez cada una.
- Muchos de los verbos que aparecen con significado léxico no se pueden considerar estrictamente vocabulario de la cerámica ya que se pueden utilizar también en otros ámbitos sin que su significado varíe. Sin embargo, su estudio o consideración sí se pueden considerar interesantes de cara a la creación de fraseología que puede ser muy útil para los redactores, traductores y revisores de textos técnico científicos de la cerámica. Es el caso de verbos como *formar*, *contrarrestar*, *producir*, *aplicar*, *condicionar*, *existir*, *obtener*, *contener*, *figurar*, *utilizar*, *aumentar*, *procesar*, *basar*, *medir*, *cargar*, *dar*, *soportar*, *elegir*, *emplear*, *determinar*, *bajar*, *permitir*, *valorar*. Como ejemplo de este uso específico, véase el siguiente contexto del verbo *contrarrestar*: "El aumento de temperatura, no obstante, tiende a contrarrestar el efecto de la sílice sobre la viscosidad".
- Otros verbos sí son específicos de la cerámica: *esmalzar*, *mezclar*, *preparar*, *solar*, *cocer*. Por último, unos pocos podrían entrar a formar parte de una stoplist ya que su contenido es más gramatical que léxico. Se trataría de verbos como *ser*, *estar*, *haber*, *hacer*, *deber*.
- Los sustantivos, son mayoritariamente específicos de la cerámica: *baldosa*, *cocción*, *óxido*, *horno*, *resistencia*, *cerámica*, *capa*, *pieza*, *composición* (química), *arcilla*, *fabricación*, *presión*. Aunque algunos también se pueden considerar sustantivos del lenguaje general: *material*, *superficie*, *tipo*, *tabla*, *sistema*, *aire*, *característica*.
- Muchos de estos sustantivos ayudan a localizar otros términos complejos que lo tienen como base. Por ejemplo, al buscar las concordancias de *presión* nos aparecen los términos complejos: *presión dinámica*, *presión estática*, *presión hidrostática*, *presión de conformación*, *presión ordinaria*, etc.

C Concord - [PRESION: 1519 entries (sort: 5L,5R)]

C File View Settings Window Help

Text Excerpt	Frequency
...férico la variación lineal entre log ((Fe2+)/(Fe3+)) y el inverso de la presión parcial de oxígeno se mantiene hasta un valor límite de po...	3 272
...ctos 74 30 - Carga - Conductos a presión y en depresión 76 31 - Presión estática, presión dinámica, presión total 78 32 - Pérdidas	724
...ción, humedad de los polvos atomizados,) y en particular a una presión de formación de sólo 250 bar, las diferentes pastas experi...	10 054
...ples factores (ciclo de cocción, dimensión y espesor de la baldosa, presión de conformación, tipo de esmalte utilizado) * Formular co	3 221
...or variables de proceso como humedad y granulometría del polvo y presión real de prensado y por la relación "plásticos/desgrasantes	2 641
...s del granulado (humedad y granulometría), estas variables son: - Presión de bombeo de la barbotina. Afecta a la producción del ato	9 539
...o, indicará evidentemente la diferencia entre h1, y h2 es decir, una presión ha correspondiente tan sólo a la fuerza viva del gas: 4) h	8 524
...errumpir la presión, el volumen crece poco o nada, y la fuerza de la presión es sustituida por las de tensión superficial en la cara exter	16 979
...or. Por medio de la temperatura del generador y la isobara de alta presión se obtiene la concentración de la solución pobre en refriger	10 929
...la aplicación son distancia entre la pistola y la línea de esmaltado, presión y caudal del aire, ángulo del abanico y tipos de boquillas	15 444
...que la curvatura depende del prensado, y más concretamente de la presión, y de la compactación de los polvos en grado diferente entr	1E
...na proporción considerable de agua y otros compuestos volátiles a presión ordinaria, pero que a las grandes presiones del interior de l	5 964
...corresponde un aumento de la cantidad de vapor, y por tanto de la presión de evaporación, lo que ocasiona el cierre de la válvula y un	10 489
...odo que el caudal en peso es superior), y sabemos además que la presión dinámica y las pérdidas de carga son proporcionales a dic	11 409
...alizada. Algunas veces la colada es bajo presión y se dice que una presión neumática de 0,35 a 1,05 kg/cm2 sobre la barbotina abrevi	1 237
...a característica intrínseca a algún óxido y viene relacionada con su presión de vapor. El óxido de plomo, el anhídrido bórico y los álcal	21 099
...do, porque hay menos adherencia a la matriz y, con los blandos, la presión se distribuye más uniformemente. Las presiones en la mat	11 589
...ón, radiación o conducción y, normalmente, coinciden. Figura 1. Presión del vapor saturado del agua a diferentes temperaturas. L	13 794
...parte, a la menor porosidad que queda en la pieza y, por otra, a la presión de los gases ocluidos en los poros cerrados. En la Figura	27 209
...n recircular parcialmente los vapores ya comprimidos hacia la baja presión (Fig. n° 10 18), este sistema presenta dos inconvenientes,	1 999
...que el pagado de la pieza al molde de yeso hacia el moldeado por presión poco práctico. Reemplazando el yeso como material del m	15 089
...binomio de dilatación 1 + at y del peso específico $\gamma_t = \gamma_0 / (1 + at)$ (a la presión atmosférica) para las diversas temperaturas t desde 0° has	2 801
...y producir una porosidad algo mayor en la zona de la esquina. La presión a lo largo de la superficie de prensado será más baja cerca	13 109
...onamiento * La carencia de separación entre zonas de alta y baja presión, por lo que al arar el compresor se produce la igualación d	13 459

Figura 2. Concordancias del término "presión"

- Han aparecido **otras palabras** que se podrían considerar también palabras gramaticales, como mayor y gran. Estas palabras deberían pasar a aumentar el listado de la stoplist en el futuro.
- Han aparecido también con gran frecuencia la abreviatura *fig* y la fórmula química 'CaO'. En el futuro habría que pensar en el tratamiento de estas expresiones quizás en un fichero de stoplist para evitar que aparezcan en el listado de términos.

4. Conclusiones y trabajos futuros

La herramienta WordList de WordSmith se ha mostrado útil y eficaz al extraer terminología específica en el ámbito de la cerámica industrial. Para ello, sin embargo, no basta con disponer de la propia herramienta, sino que ha sido necesario elaborar recursos específicos para el tratamiento del corpus en español. También se ha podido apreciar la necesidad de elaborar stoplist y ficheros de lemas más precisos, incorporando los nuevos elementos que van surgiendo al hacer el análisis.

El listado de términos resultante es bueno por sí mismo y también porque constituyen sondas para extraer nuevos términos complejos que contienen esos términos simples como base léxica.

Algunas palabras de las que han aparecido en las primeras cien, podrían pasar a formar parte de una stoplist (verbos carentes de significado léxico, abreviaturas de uso general y otras palabras gramaticales que no habían sido recogidas al principio). Con ello conseguiríamos un listado aún más depurado de términos.

Los recursos que se han generado han permitido crear un listado depurado de términos de la cerámica. Ahora bien, además, estos recursos son reutilizables para el estudio de otros corpus de otros ámbitos de especialidad.

La herramienta WordList, sin embargo, ofrece pocas comodidades para el trabajo cotidiano del traductor. Frente a otras herramientas como puede ser la herramienta ExtraTerm que forma parte de un paquete de traducción asistida, la herramienta WordSmith resulta complicada y poco amigable. Tampoco ofrece posibilidades de integración con otros programas de traducción asistida.

Sería deseable en el futuro poder integrar las ventajas y la eficacia de una herramienta lexicográfica con las ventajas de integración con la traducción que ofrecen otros programas.

En lo que respecta a nuestra investigación, se han llevado algunos pasos para comprobar la eficacia de una herramienta como ExtraTerm con resultados de fiabilidad muy bajos, aunque todavía provisionales. También se ha probado la herramienta KeyWord, de WordSmith, con muy buenos resultados. Esta herramienta trabaja con listas de palabras extraídas de dos corpus: un corpus de especialidad y un corpus de referencia (para más información puede verse Berber Sardinha 1999). Mientras no desarrollemos u obtengamos un corpus de referencia de calidad no podremos probarla plenamente.

Bibliografía

- Alcina, A. y Pruñonosa Tomás, M. (1992). "Uso del ordenador para el estudio terminológico previo a la traducción de un texto". En *Actes del I Congrés Internacional sobre Traducció* M. Edo Julià (ed., 103-112. Bellaterra: Universitat Autònoma de Barcelona.
- Berber Sardinha, T. (1999). "Usando WordSmith Tools na investigação da linguagem". *DIRECT Papers*, 40, Faber, P. y Jiménez, C. eds.) (2002). *Investigar en Terminología*. Granada: Comares.
- López Rodríguez, C. I. (2001). "Detección automática de la cohesión léxica en textos sobre oncología: aplicaciones a la traducción". En *Traducción y nuevas tecnologías. Herramientas auxiliares del traductor* C. Valero Garcés y I. De La Cruz Cabanillas eds.), 327-337. Alcalá de Henares: Universidad de Alcalá.
- Sánchez-Gijón, P. (2005). *L'ús de corpus en la traducció especialitzada: Compilació de corpus ad hoc o extracció de recursos terminològics*. Gerona: Documenta Universitaria.
- Scott, M. *WordSmith*, University of Oxford.
- Soler Puertes, V., Alcina Caudet, M. A. y Estellés Palanca, A. (2005). El trabajo en equipo con estudiantes para la elaboración de un corpus lingüístico electrónico. En *X Jornades de Traducció i Interpretació a Vic: Tecnologies a l'abast*, Vic.