ANALISIS AUTOMATICO DE DOCUMENTACION TECNICA INFORMATICA

Félix de Moya
José Manuel Muñoz
Pedro Hípola

Moya, F.; Hípola, P. «Análisis automático de documentación técnica informática». *V Congreso AESLA*. Pamplona, 1987.
http://www.ugr.es/~phipola/Analisis_automatico_de_documentacion_tecnica_informatica.pdf

La selección del vocabulario específico para la enseñanza de lenguajes especializados ha de estar fundamentada en estudios lexicográficos rigurosos, para los cuales es posible hoy da contar con las técnicas usuales de la informática documental.

En esta comunicación se presenta un ejemplo práctico de una metodología, con la que los autores vienen trabajando desde hace algún tiempo, encaminada al análisis de los aspectos léxicos de los textos que se emplean en la enseñanza de la lengua inglesa para fines específicos (E.S.P.). El trabajo parte de la creación de una base de datos documental, a partir de la cual se realiza automáticamente el análisis estadístico de frecuencias que permita establecer cuáles son los términos específicos ms utilizados en la documentación informática. Para ello se ha procesado un corpus textual de más de doscientas mil palabras.

En estas últimas décadas se han realizado muchos esfuerzos para hacer avanzar el proyecto de la automatización en el tratamiento de material lingüístico. Para ello se ha contado con las sucesivas aportaciones que han ofrecido las ciencias del lenguaje y la tecnología de los ordenadores, así como la experiencia acumulada en centros de procesamiento documental automatizado.

Los logros obtenidos hasta ahora pueden ser aprovechados para diferentes aplicaciones. En estas páginas nos vamos a referir a la interacción que existe entre el uso de técnicas informáticas documentarias y el establecimiento de vocabularios correspondientes a corpus textuales específicos. Expondremos de forma breve el proceso que hemos seguido para la preparación de un vocabulario de términos técnicos informáticos, elaborado a partir de textos ingleses que se han seleccionado para la enseñanza de esta lengua en el marco del proyecto didáctico que suele englobarse bajo la expresión E.S.P. (English for Specifical Purposes).

La selección del vocabulario que se va a emplear en la enseñanza de una lengua extranjera ha sido y es tema de constante estudio en el campo de la lingüística aplicada. Se trata de un ámbito de trabajo permanentemente abierto a la investigación y al análisis, no sólo por la propia naturaleza cambiante del lenguaje, que hace del vocabulario un cuerpo en continua evolución, sino también por la irrupción de fenómenos nuevos que engendran nuevas demandas expresivas. Puesto que el léxico de una lengua constituye de hecho un corpus inabarcable, no ya para un individuo, sino

para toda una comunidad lingüística, se hace necesario determinar los criterios en torno a los cuales se basar la selección del léxico que se va a hacer aprender.

Ya desde los primeros estudios realizados, se señalaba el criterio de finalidad para establecer los vocabularios, basándose en una interpretación instrumental del lenguaje. El comit del Interim Report on Vocabulary Selection (1) hablaba de nueve posibles "purposes" en el aprendizaje. Pues bien, entre los fines que se inventariaron no se haca mención al lenguaje de la Ciencia y la Tecnología. Tampoco el comité del "français foundamental" contemplaba el lenguaje científico y técnico como uno de los campos en el que una selección de vocabulario fuera necesaria. Sin embargo, el lenguaje de la ciencia y de la técnica con el tiempo se ha ido convirtiendo, junto con el lenguaje comercial, en el que ms demanda social suscita entre los estudiantes de una segunda lengua. Avances recientes en el campo de la investigación científica y tecnológica han hecho surgir nuevas orientaciones para la enseñanza de idiomas, que, bajo la denominación general de E.S.P., se centran no en la lengua general común, sino en tipos de lengua específicos de acuerdo con las necesidades reales del aprendiz.

En este contexto está claro que resulta imprescindible determinar la estructuración de vocabularios específicos para cada actividad. La elección de las palabras puede hacerse de acuerdo con los criterios ms o menos subjetivos del profesor, que decide enseñar esto y no aquello o establecer un orden en lo que enseña; pero existe también, y se ha venido concibiendo desde el principio, la idea de que sería deseable que la selección se haga a partir de rigurosos estudios estadísticos de frecuencias léxicas. Las aportaciones de West y Lorge (1) son quizá el ejemplo más significativo de la aplicación de este principio. Su incidencia en el ámbito de la didáctica ha sido decisiva.

El trabajo que presentamos pretende ser una contribución ms a la ya larga lista de estudios que buscan encontrar soluciones al siempre complejo problema de la selección del vocabulario. El campo específico en el que se ha desarrollado es el del inglés en textos técnicos relacionados con la informática (English for computer science/English for science and technology), y la metodología aplicada para su elaboración -que de momento no incluye rutinas de análisis contextual- ha seguido las siguientes etapas:

1. selección del corpus de lengua sobre el que realizar el estudio,
2. depuración automática,
3. recuento de frecuencias léxicas,
4. evaluación de resultados.

Para la realizacin de la primera etapa se seleccionaron textos en inglés de ayuda "on line" para diversos sistemas operativos, así como para programas de aplicaciones de amplia difusión: procesadores de texto, bases de datos, lenguajes de alto nivel, etc. En total, algo ms de doscientas cincuenta mil palabras (tokens).

En primer lugar, se crea la correspondiente base de datos, cuyos campos estuvieron formados por secuencias de texto con una longitud no superior a doscientos cincuenta caracteres, con el fin de acomodar los registros lógicos a los registros físicos gestionados por el sistema utilizado, de modo que la depuración formal del corpus fuera ms rápida. Este proceso de depuración se lleva a cabo de forma automática por un módulo de programa codificado a tal efecto. Así fueron analizados todos los caracteres que previsiblemente pudieran perturbar el recuento automático de las distintas formas léxicas. En algunos casos dichos caracteres -la mayor parte de los

signos de puntuación- quedaron eliminados. En otras ocasiones -algunos caracteres que se emplean para abreviaturas, acrónimos, palabras compuestas- se respetaron o sustituyeron por otros. Todos los caracteres alfabéticos que aparecían en minúscula fueron transformados a mayúscula para que quedaran homogeneizadas las formas léxicas con independencia de su situación en el texto. Por último, se suprimieron algunos elementos irrelevantes, tales como fórmulas matemáticas, etc., quedando como resultado un volumen algo inferior a doscientas mil formas léxicas. Realmente, no interesaba que el corpus fuera de un tamaño mayor, para que no se desvirtuara la comparación que se iba a realizar con otro corpus de carácter general, el de Kucera y Francis (3). Ha de tenerse en cuenta que, al ocuparnos nosotros de un tipo determinado de lenguaje la comparación habrá de hacerse con respecto al número de palabras procesadas para cada uno de las quince reas temáticas en que Kucera y Francis dividen su corpus y cuya magnitud no excede en ningún caso las cien mil palabras.

El recuento de frecuencias léxicas fue realizado, creando una segunda base de datos en la que se introdujeron como entrada principal cada una de las formas léxicas. Aparte, en otro campo fue acumulado el número de ocurrencias de las formas. El resultado fue un total de ms de cinco mil registros.

A continuación se editó un tercer campo, en el que se consigna la frecuencia relativa de aparición de cada forma en el corpus general (Kucera); y otros ms en los que se dispusieron informaciones de carácter lingüístico:

-indicador de lexical canónico s/no,

-en su caso, referencia al término lexical canónico correspondiente, normalmente el infinitivo (4),

-palabra vaca s/no (5),

-anotaciones especiales, si procede: abreviaturas, etc.

Llegados a este punto, se considera preparada la herramienta básica a partir de la cual se podrá constituir el léxico específico. Esto será posible a base de cubrir diferentes etapas, la primera de las cuales fue realizar la acumulación de frecuencias: suma de todas las frecuencias absolutas de las formas flexionadas en la frecuencia del término lexical canónico correspondiente, tanto en el campo que corresponda al corpus específico como al del general. Una vez conocidos estos dos tipos de frecuencias absolutas, haba que proceder a compararlos, pues, como es sabido, la simple aparición de un determinado vocablo un número elevado de veces en el texto no es dato suficiente para considerarlo término específico. Esta decisión debe tomarse tras un detenido análisis estadístico. Para solucionar el problema de cómo establecer la especificidad de un determinado lexical canónico se había decidido -como ya hemos adelantado- emplear los estudios de frecuencias de Kucera y Francis como término de la comparación, por varios motivos: en primer lugar, porque la metodología del trabajo de Kucera y Francis nos parecía acorde con la empleada en nuestro estudio; en segundo lugar, por su general aceptación como uno de los estudios ms completos en este campo. Esta elección tena también otros extremos que deban ser considerados: la obra de Kucera y Francis se refiere al inglés americano y está realizada exclusivamente sobre textos escritos. Por tanto no resulta un reflejo exacto de lo que sería el inglés general. Estos extremos no nos parecieron problemas de gran entidad, ya que, en definitiva, nuestro análisis se basaba también en textos escritos, muchos de ellos

producidos en E.E.U.U. Por otro lado la división en quince categorías de textos de acuerdo con su temática ofrecía la posibilidad de considerar nuestro estudio como complementario del de Kucera y Francis.

Para intentar establecer la especificidad de un término recurrimos, pues, a criterios de tipo estadístico basados en la frecuencia relativa de cada lexical canónico, su distribución en nuestro corpus y la comparación con otros corpus, en este caso con el de Kucera. Para ello contábamos con las técnicas de análisis usuales en lingüística estadística, como el test de Pearson, la ley de Zipf, etc.

Se solicitó al sistema un listado de los lexicales canónicos con una frecuencia relativa, dentro de nuestro corpus específico, superior a la frecuencia relativa del corpus general. Como es natural, aparecieron demasiados términos, unos que podrán ser considerados muy específicos de la materia en cuestión, pero otros que evidentemente no lo eran. En una segunda aproximación se requirió un listado de estos mismos términos, con exclusión de las palabras caracterizadas como vacas y las de frecuencia absoluta 1 ("hapax legomene"). Se produjo entonces una lista, jerarquizada de acuerdo con un orden decreciente de ocurrencias, que, a nuestro juicio, puede considerarse en general suficientemente válida y resultar un auxiliar relevante para proyectos de E.S.P. Añadamos que la relación de términos que se obtuvo contribuyó a poner de manifiesto que el corpus general de Kucera y Francis, a pesar de ser relativamente reciente, al ser puesto en comparación con documentación sobre un tema -la informática- de un denso crecimiento terminológico en los últimos años, revela importantes ausencias, especialmente en lo que se refiere a determinadas siglas.

En cualquier caso, y como conclusión, queremos dejar constancia de que, una vez ms, se demuestra la validez de métodos como los que se han empleado en este estudio, cada da ms asequibles para todos. El sistema permite determinar con bastante precisión los grados de especificidad del vocabulario de un área temática determinada. Por otra parte, a medida que la base de datos léxica aumenta, el trabajo de actualización de los léxicos específicos es ms sencillo. Todo apunta hacia la posibilidad de que mantener al da glosarios específicos sea algo que se realice cada vez ms a travs de procedimientos automáticos.

\ctr\REFERENCIAS BIBLIOGRAFICAS

(1) Publicado en London el ao 1936.

(2) WEST, M., A General Service List of English Words, London, 1953. Cfr. también el trabajo de Mckey, Language Teaching analysis.

(3) KUCERA, H. y FRANCIS, W. N., Computational Analysis of Present-day American English_, Rhode Island, 1967.

(4) Al llegar a este punto interesa señalar que nosotros no identificamos término lexical canónico con lema, sino que utilizamos esta denominación para referirnos a lo que A. Deweze llama "forma semántica". Cfr. Informatique documentaire_, Paris, 1985, pg. 129 y ss.

(5) Sólo se caracterizaron como vacas unidades léxicas que son consideradas generalmente las más gramaticales, con menor carga semántica propia: elementos de relación, determinantes, etc. No se utilizó el criterio estadístico para establecer palabras "temticas", por los riesgos que para los corpus específicos puede entrañar este procedimiento.
Para ello se ha procesado un corpus textual de ms de doscientas mil palabras.